Multiple sclerosis (MS) is an immune-mediated neurological disease that causes disability and morbidity. In patients with MS, the accumulation of lesions in the white matter of the brain is associated with disease progression and worse clinical outcomes. In the first part of the dissertation, we present methodology to study the brain anatomy between patients with MS and controls. A nonparametric testing procedure is proposed for testing the null hypothesis that two samples of curves observed at discrete grids and with noise have the same underlying distribution. We propose to decompose the curves using functional principal component analysis of an appropriate mixture process, which we refer to as marginal functional principal component analysis. This approach reduces the dimension of the testing problem in a way that enables the use of traditional nonparametric univariate testing procedures. The procedure is computationally efficient and accommodates different sampling designs. Numerical studies are presented to validate the size and power properties of the test in many realistic scenarios. In these cases, the proposed test is more powerful than its primary competitor. The proposed methodology is illustrated on a state-of-the-art diffusion tensor imaging study, where the objective is to compare white matter tract profiles in healthy individuals and MS patients.

In the second part of the thesis, we present methods to study the behavior of MS in the white matter of the brain. Breakdown of the blood-brain barrier in newer lesions is indicative of more active disease-related processes and is a primary outcome considered in clinical trials of treatments for MS. Such abnormalities in active MS lesions are evaluated in vivo using contrast-enhanced structural magnetic resonance imaging (MRI), during which patients receive an intravenous infusion of a costly magnetic contrast agent. In some instances, the contrast agents can have toxic effects. Recently, local image regression techniques have been shown to have modest performance for assessing the integrity of the blood-brain barrier based on imaging without contrast agents. These models have centered on the problem of cross-sectional classification in which patients are imaged at a single study visit and pre-contrast images are used to predict post-contrast imaging. In this paper, we extend these methods to incorporate historical imaging information, and we find the proposed model to exhibit improved performance. We further develop scan-stratified case-control sampling techniques that reduce the computational burden of local image regression models while respecting the low proportion of the brain that exhibits abnormal vascular permeability.

In the third part of this thesis, we present methods to evaluate tissue damage in patients with MS. We propose a lag functional linear model to predict a functional response using multi-
ple functional predictors observed at discrete grids with noise. Two procedures are proposed to estimate the regression parameter functions; 1) a semi-local smoothing approach using generalized cross-validation; and 2) a global smoothing approach using a restricted maximum likelihood framework. Numerical studies are presented to analyze predictive accuracy in many realistic scenarios. We find that the global smoothing approach results in higher predictive accuracy than the semi-local approach. The methods are employed to estimate a measure of tissue damage in patients with MS. In patients with MS, the myelin sheaths around the axons of the neurons in the brain and spinal cord are damaged. The model facilitates the use of commonly acquired imaging modalities to estimate a measure of tissue damage within lesions. The proposed model outperforms the cross-sectional models that do not account for temporal patterns of lesional development and repair.
Statistical Methods for Magnetic Resonance Image Analysis with Applications to Multiple Sclerosis

by
Gina-Maria Pomann

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

Statistics

Raleigh, North Carolina
2015

APPROVED BY:

Marie Davidian

Montserrat Fuentes

Sujit Ghosh

Russell Taki Shinohara

Ana-Maria Staicu
Chair of Advisory Committee
DEDICATION

For Edgar, Mom, Dad, Olivia, and my Grandparents.
BIOGRAPHY

Gina-Maria Pomann's passion is to work on projects that have potential to positively impact society. Throughout her academic career, she has worked to promote diversity in the sciences and to conduct applied research with broad impacts. Her graduate studies were supported by the National Heart Lung and Blood Institute (NHLBI) Traineeship, the National Science Foundation Graduate Research Fellowship, and the AT&T Labs Graduate Research Fellowship.
ACKNOWLEDGEMENTS

There have been so many people that made it possible for me to write this thesis, and it is my pleasure to acknowledge them. First, I wish to express my deepest gratitude to my Advisor, Ana-Maria Staicu for her support and guidance throughout my graduate career. This research would not have been possible without her guidance. For collaborating and providing invaluable insight necessary to complete this work I would like to sincerely thank, Russell Taki Shinohara, Sujit Ghosh, Elizabeth Sweeney, Danny Reich, and Blake Dewey. I would like to thank the members of PennSIVE group for sharing my passion for medical imaging and helping me conduct research in this area. In particular, I would like to thank Taki Shinohara for teaching me so much about MRI and Elizabeth Sweeney creating so many enlightening images used to present our work. I am exceedingly grateful to my mentors, Marie Davidian, Montserrat Fuentes, and Sujit Ghosh for serving on my committee, helping me to focus my research interests, and guiding me through the process. I also thank Wesley Snyder for serving as the graduate school representative on my committee.

With gratitude, I am pleased to acknowledge all of the funding sources that supported my research: the National Science Foundation under Grant No. DGE-0946818, the AT&T Graduate Research Fellowship, the National Heart Lung and Blood Institute (NHLBI) Traineeship, and the NIH grant RO1 NS085211 from the National Institute of Neurological Disorders and Stroke (NINDS). The studies were supported in part by the Intramural Research Program of NINDS. This work represents the opinions of the researchers and not necessarily that of the granting organizations.

I thank all of the faculty and my professors in the North Carolina State University (NCSU) Statistics department for providing a superb education and support structure. I am exceedingly grateful for my fellow graduate students in this department. In particular, I can’t imagine graduate school without Kristin Linn, Kasturi Talapatra, Beth Ann Tidemann-Miller, and Alana Unfried who are now my lifelong friends. I also thank all of the staff at NCSU. In particular, Alison McCoy, Terry Byron, and Chris Waddell are so kind and always know how to solve any problem. I sincerely thank Marcia Gumpertz, David Shafer, Mark Newmiller, and Sabina Vermeulen who made it possible for me pursue my passion for promoting the advancement of students with diverse backgrounds.

The outstanding support of my mentors paved the way for me to be apart of the NCSU Department of Statistics. In high school, Ilene Kesselhaut, Sherri Horn and Arlene Bernstein helped me gain the confidence I needed to attend college. Without my professors at Middlesex County College I would never have pursued a degree in Mathematics. In particular, Richard Klein and George Popel always went the extra mile to show me how this field can change our
world. For this, I am profoundly grateful.

The Mathematical Sciences Research Institute - Undergraduate Program (MSRI-UP) was a vital part of my journey. The organizers, and in particular my mentors, Ricardo Cortez and Juan Meza have provided an inconceivable level of support and guidance over the last 8 years. I attribute the majority of my academic accomplishments during this time to what I learned as a participant in this program.

I am grateful to the faculty and staff in the Mathematics and Statistics department at The College of New Jersey. In particular, Aigli Papantonopoulou for encouraging me to attend the MSRI-UP and then graduate school. Leona Harris has been the greatest teacher, friend, and mentor I could ever imagine having in my life. To Dr. Harris, I express my deepest gratitude.

I thank AT&T Labs for providing financial support as well as invaluable guidance throughout my graduate career. In particular, my research advisor, Tamraparni Dasu who guided me through my first graduate level research experience and has been a remarkable mentor throughout my studies; Robert Bell, who provided extraordinary support and guidance through my studies; and Chris Volinsky who helped me shape my goals and pursue my passions.

Most importantly, I thank my friends and family for their unwavering support. I have the deepest gratitude for Edgar who has helped me gain so much insight and has fueled much of my passion for research. For an immeasurable set of love, coding, proofs, editing, and philosophical debates about ‘Gina Math’ I cannot thank you enough. I am grateful to my father for his support and pride. I thank my sister for always having faith in my abilities and for joining me on this journey. My Uncle George and Auntie Sue guided me towards my passion for applications in Neurology. I am so grateful for the support of my grandparents, Emile Levy and Evelyn Pomann. In particular, my grandmother is one of my best friends and has always conveyed the importance of education. My Aunt Esther, Uncle Yigal, and all of my cousins have provided so much love through this process. My pets are also my family and always make me smile providing good company throughout my studies.

I would like to express the most profound gratitude to my mother for conveying the teaching, “My work is to find my world and with all of my heart give myself to it.” She has shown me how to stay true to my beliefs and overcome all obstacles. Her support and guidance made it possible for me to write this thesis. I thought I could, I thought I could, I thought I could...
TABLE OF CONTENTS

LIST OF TABLES .......................................................... viii

LIST OF FIGURES ........................................................ x

Chapter 1 Introduction .................................................. 1
  1.1 Motivation .......................................................... 1
  1.2 Functional Data ................................................... 2
    1.2.1 Smoothing Using Basis Function Expansions ............... 3
    1.2.2 Smoothing Using Roughness Penalties ..................... 4
    1.2.3 Functional Principal Components Analysis ............... 5

Chapter 2 A Two-Sample Distribution-Free Test for Functional Data with
  Applications to Diffusion Tensor Imaging ......................... 8
  2.1 Background and Motivation .................................... 8
  2.2 Two-Sample Testing for Functional Data ...................... 10
  2.3 Extension to Practical Situations ............................ 13
    2.3.1 Dense design ............................................. 14
    2.3.2 Sparse design ............................................ 14
  2.4 Simulation Studies ............................................. 15
    2.4.1 Type One Error and Power Performance ................... 16
    2.4.2 Comparison with an Alternative Approach ................ 19
  2.5 Diffusion Tensor Image Data Analysis ......................... 21

Chapter 3 Scan-Stratified Case Control Sampling for Modeling Blood-Brain
  Barrier Integrity in Multiple Sclerosis .......................... 25
  3.1 Background and Motivation .................................... 25
  3.2 NINDS Data ..................................................... 27
  3.3 Logistic Enhancement Models ................................... 28
  3.4 Sampling on 3D Magnetic Resonance Images .................... 29
  3.5 Enhancement Prediction Results ................................ 31

Chapter 4 A Lag Functional Linear Model for Prediction of Magnetization
  Transfer Ratio in Multiple Sclerosis Lesions .................... 35
  4.1 Introduction .................................................... 35
  4.2 Methodology .................................................... 37
    4.2.1 PW Model Representation ................................ 38
    4.2.2 GB Model Representation ................................ 44
  4.3 Simulation Studies ............................................. 46
    4.3.1 Simulation Results ....................................... 48
  4.4 Prediction of Magnetization Transfer Ratio within Lesions of MS Patients .. 50
    4.4.1 Data Processing .......................................... 51
    4.4.2 Model Implementation and Results ....................... 52
LIST OF TABLES

Table 2.1  Estimated Type I Error rate of FAD Test, based on 5000 replications for several threshold values, \( \tau \), leading to different estimates of the truncation parameter \( K \). .................................................. 18

Table 3.1  Partial AUC estimates and corresponding 95% confidence intervals (CI) for curves in figures 3.1 and 3.2. .................................................. 32

Table 3.2  Coefficient estimates when model 1 (left) and model 2 (right) are fit using all voxels in the corresponding sets. .................................................. 34

Table 4.1  Prediction Error of PW Method - Setting (I), Average Error Variance = 0.012 49
Table 4.2  Prediction Error of GB Method - Setting (I), Average Error Variance = 0.012. 50
Table 4.3  Prediction Error of PW Method - Setting (II), Average Error Variance = 0.012. .................................................. 50
Table 4.4  Prediction Error of GB Method - Setting (II), Average Error Variance = 0.012. .................................................. 50
Table 4.5  Prediction error from 10-fold cross validation on MTR data with two predictors. Integrated mean squared error (MSE) times \( 10^3 \) along with the average \( MSE/s_0^2 \), where \( s_0^2 \) is the estimated measurement error in the observed MTR voxel trajectories, for all models and estimation approaches presented in the paper. .................................................. 54

Table B.1  Prediction Error for PW - Setting (I), One Predictor, eSNR = 1, Average Error Variance = 0.043 .................................................. 75
Table B.2  Prediction Error for GB - Setting (I), One Predictor, eSNR = 1 Average Error Variance = 0.043 .................................................. 75
Table B.3  Prediction Error for PW - Setting (I), One Predictor, SNR = 5 , Average Error Variance = 9e-03 .................................................. 75
Table B.4  Prediction Error for GB - Setting (I), One Predictor, eSNR = 5 , Average Error Variance = 9e-03 .................................................. 75
Table B.5  Prediction Error for PW - Setting (I), Two Predictors, eSNR = 1 , Average Error Variance = 0.058 .................................................. 76
Table B.6  Prediction Error for GB - Setting (I), Two Predictors, eSNR = 1 , Average Error Variance = 0.058 .................................................. 76
Table B.7  Prediction Error for PW - Setting (II), One Predictor, SNR = 1 , Average Error Variance = 0.043 .................................................. 77
Table B.8  Prediction Error for GB - Setting (II), One Predictor, eSNR = 1 , Average Error Variance = 0.043 .................................................. 77
Table B.9  Prediction Error for PW - Setting (II), One Predictor, eSNR = 5, Average Error Variance = 9e-03 .................................................. 78
Table B.10 Prediction Error for GB - Setting (II), One Predictor, eSNR = 5, Average Error Variance = 9e-03 .................................................. 78
Table B.11 Prediction Error for PW - Setting (II), Two Predictors, eSNR = 1 , Average Error Variance = 0.058 .................................................. 79
Table B.12 Prediction Error for GB - Setting (II), Two Predictors, eSNR = 1, Average Error Variance = 0.058

Table B.13 Prediction Error for PW - Setting (I), One Predictor, idd Error, eSNR = 1, Average Error Variance = 0.043

Table B.14 Prediction Error for GB - Setting (I), One Predictor, idd Error, eSNR = 1, Average Error Variance = 0.043

Table B.15 Prediction Error for PW - Setting (I), One Predictor, idd Error, eSNR = 5, Average Error Variance = 9e-03

Table B.16 Prediction Error for GB - Setting (I), One Predictor, idd Error, eSNR = 5, Average Error Variance = 9e-03

Table B.17 Prediction Error for PW - Setting (I), Two Predictors, idd Error, eSNR = 1, Average Error Variance = 0.058

Table B.18 Prediction Error for GB - Setting (I), Two Predictors, idd Error, eSNR = 1, Average Error Variance = 0.058

Table B.19 Prediction Error for PW - Setting (II), One Predictor, idd Error, eSNR = 1, Average Error Variance = 0.043

Table B.20 Prediction Error for GB - Setting (II), One Predictor, idd Error, eSNR = 1, Average Error Variance = 0.043

Table B.21 Prediction Error for PW - Setting (II), One Predictor, idd Error, eSNR = 5, Average Error Variance = 9e-03

Table B.22 Prediction Error for GB - Setting (II), One Predictor, idd Error, eSNR = 5, Average Error Variance = 9e-03

Table B.23 Prediction Error for PW - Setting (II), Two Predictors, idd Error, eSNR = 1, Average Error Variance = 0.058

Table B.24 Prediction Error for GB - Setting (II), Two Predictors, idd Error, eSNR = 1, Average Error Variance = 0.058

Table B.25 Prediction error for MTR data analysis with one predictor from 10-fold cross validation. Integrated mean squared error (MSE) times \((10^2)\) along with the average \(\text{MSE}/s_0^2\), where \(s_0^2\) is the estimated measurement error in the observed MTR voxel trajectories.
LIST OF FIGURES

Figure 2.1 Empirical power curves under simulation setting A (leftmost panels), B (middle panels) and C (rightmost panels). The top panels consist of the case in which both sets of curves have equal sample sizes and the bottom panels correspond to unequal sample sizes as displayed in the legends. The overall sample size \( n = n_1 + n_2 \) varies from \( n = 200 \) to \( n = 1000 \). The maximum standard error is 0.007. .................................................. 18

Figure 2.2 Estimated power curves for the FAD (solid line) and CVM (dashed) for equal sample sizes \( n_1 = n_2 \) varying from 15, 25 to \( n = 50 \) and corresponding to \( m_1 = m_2 = 20 \) (black-dot) and \( m_1 = m_2 = 100 \) (blue-square). Level of significance is \( \alpha = 0.05 \). 500 MC replications are used. ................................. 20

Figure 2.3 Top: Fractional anisotropy for cases and controls with the pointwise mean in red. Bottom: Parallel diffusivity for cases and controls with the pointwise mean in red. .................................................. 21

Figure 2.4 Parallel Diffusivity (LO). Top: First three estimated eigenfunctions that explain 91% of the variation in the combined data set. Five eigenfunctions explain over 95% of the marginal variation in the data. Bottom: Box plots of the first three estimated basis coefficients. ................................. 22

Figure 2.5 Fractional Anisotropy (FA). Top: First four estimated eigenfunctions that explain over 90% of the total variation in the combined FA data set. Six eigenfunctions explain over 95% of the marginal variation in the data. Bottom: Box plots of the first four estimated marginal basis coefficients. ................................. 23

Figure 3.1 Partial ROC curves comparing Models (1) and (2) (blue line (circles) and solid red line (squares) respectively) for; A-All voxels; B-Top 1% FLAIR sub-sampling; and C-SSCC sub-sampling. The dashed line in all curves corresponds to having TPR=FPR. .................................................. 32

Figure 3.2 Partial ROCs curves from using the three data sets under study to fit model (1) (A) and model (2) (B). .................................................. 32

Figure 3.3 Top row (left to right): T1w pre-contrast; T2w pre-contrast; FLAIR postcontrast; T1w post-contrast. Bottom row (left to right): SuBLIME Map; Model 1 prediction map using FLAIR thresholding (top 1%) to fit and predicting on the full brain, as per Shinohara et al.(2012); Model 2 prediction map with case control sampling for fitting and predicting on the full data set; RGS mask. .................................................. 33

Figure 4.1 From top to bottom, axial slices of an MS lesion in the FLAIR (A), T1w (B), and MTR (C) sequences. The MS patient is imaged over time in weeks from left to right. Time zero corresponds to lesion incidence and the lesion on each sequence is denoted at time zero with a red arrow. On the MTR and T1w images the lesion is the hypointense region; on the FLAIR the lesion is hyper intense. .................................................. 55
Figure 4.2  (a) T1w voxel intensities for a subset of the lesion voxels in one patient; (b) corresponding smoothed T1w voxel intensities using 20 basis functions to smooth all of the data; (c) corresponding FLAIR voxel intensities; (d) corresponding smoothed FLAIR voxel intensities using 20 basis functions to smooth all of the data. 56

Figure 4.3  (a) MTR voxel intensities for a subset of the lesion voxels in one patient; (b) smoothed MTR voxel intensities using 20 basis functions to smooth all of the data; (c) predicted MTR voxel intensities using the GB method; (d) predicted MTR voxel intensities using the PW method. 56

Figure B.1  Observed data under simulation settings (I) - row one and (II) - row two. 73
Figure B.2  Smoothed predictor functions using 20 basis functions. 74
Figure B.3  Predicted response functions using the PW estimation method (row one) and the GB method (row two). 74
Chapter 1

Introduction

1.1 Motivation

Multiple sclerosis (MS) is an immune-mediated neurological disease that causes morbidity and disability. In patients with MS, the myelin sheaths around the axons of the neurons in the brain and spinal cord are damaged. Lesions in the brain and spinal cord are known to characterize this disease. In patients with relapsing-remitting MS, the most common first stage of the disease (Goodin, 2014, Chapter 15), the accumulation of lesions in the white matter of the brain is associated with disease progression and worse clinical outcomes. Breakdown of the blood-brain barrier (BBB) in newer lesions is indicative of more active disease processes. Such abnormalities in active MS lesions are evaluated in vivo using contrast-enhanced structural magnetic resonance imaging (MRI). Lesions exhibiting these abnormalities are referred to as enhancing due to their hyperintense presentation on T1-weighted (T1w) images after intravenous administration of a contrast agent. The number and volume of such enhancing lesions are important for the clinical management of patients with MS and are primary outcomes of clinical trials of treatments for MS (McDonald et al., 2001; Polman et al., 2005). Conventional MRI modalities that are used to observe lesions are T1w, T2-weighted, and T2-weighted fluid attenuated inversion recovery (FLAIR) imaging.

Diffusion tensor imaging (DTI) tractography is also commonly used to study the anatomy of patients with MS. DTI is a MRI technique that measures water diffusivity in the brain and can be used to visualize the white matter tracts of the brain. These tracts are regions of the brain commonly known to be affected by MS. There are many different modalities that can be used to measure the water diffusivity. In this thesis, we focus on the parallel diffusivity (LO) and fractional anisotropy (FA). LO quantifies the magnitude of diffusivity in the direction of the tract, whereas FA represents the degree of anisotropy. FA is equal to zero if water diffuses perfectly isotropically and to one if it diffuses anisotropically, that is along a perfectly
organized direction. The scientific problem of interest addressed in Chapter 2 is whether the parallel diffusivity or fractional anisotropy along the corpus callosum, the largest white matter tract, have identical distributions in MS patients and healthy controls. For this purpose, we develop a two-sample distribution-free hypothesis test for functional data.

In Chapter 3, we shift our focus to the investigation of the breakdown of the BBB in patients with MS. The standard procedure for assessing changes in the BBB involves the comparison of pre-contrast MRIs with MRIs acquired after the intravenous infusion of a gadolinium chelate, a magnetic contrast agent. However, contrast-enhanced imaging can cost 38% more than taking an MRI without contrast (Medicare, 2014). Gadolinium-based contrast agents have also rarely been associated with kidney problems as well as allergic reactions (Moreno-Romero et al., 2007; Kay, 2008; Perazella, 2009). Thus, methodology for assessing the integrity of the BBB based on contrast-free imaging potentially has broad clinical implications. Chapter 3 presents novel methodology using standard pre-contrast MRI modalities along with a historical covariate in a voxel-level model to predict enhancement.

It has been argued that these images only moderately correlate with disability (Hawkins et al., 1990; Brex et al., 2002; Schmierer et al., 2004). Therefore, there have been a number of studies that investigate the use of images that directly quantify tissue damage such as magnetic transfer ratio (MTR) maps. It is of scientific interest to retrospectively estimate these quantitative maps using only commonly acquired MRI modalities. Furthermore, strong predictive ability may indicate that much of the information about tissue damage that is visible in these quantitative images is also contained in the commonly acquired modalities. In Chapter 4, we present a novel functional regression framework to predict MTR in lesions using commonly acquired MRI modalities. The next section provides an introduction to functional data analysis methodology employed in this thesis.

### 1.2 Functional Data

In functional data analysis (FDA) each data point is considered to be an independent realization of some underlying (stochastic) process observed with noise on a finite grid of points. We observe data, $[Y_{ij}, t_{ij} : i \in \{1, \ldots, n\} \text{ and } j \in \{1, \ldots, m_i\}]$ where $t_{ij} \in \mathcal{T}$ for some bounded and closed interval $\mathcal{T}$; for simplicity we often take $\mathcal{T} = [0, 1]$. The index $i \in \{1, \ldots, n\}$ correspond to the realization of the stochastic process, and the index $j \in \{1, \ldots, m_i\}$ is associated with the corresponding value in the domain where the process is sampled. As observed from the notation, the observation points $t_{ij}$ do not have be the same across different realizations of the process. One key assumption for FDA is that these observation points have some inherent order (such as time or location).
In FDA we commonly consider the model,
\[ Y_{ij} = X_i(t_{ij}) + \epsilon_{ij} \]  
(1.1)
where \( X_i(t) \) is a square-integrable random function over \( T \). It is commonly assumed that \( X_i(t) \) has unknown continuous mean, and continuous and positive semi-definite covariance functions. The measurement error, \{\epsilon_{ij}\}_{i,j}, is independent and identically distributed with mean zero and variance \( \sigma^2 \). In FDA, the goal is often to make inference about the smooth underlying process \( X_i(t) \). In order to do this, many techniques can be employed. We provide a brief overview of some of these methods in the following sections.

The number of observations on each curve, \( m_i \), is either large or small. When this number is large, we refer to the functional data as ‘dense’. Whereas when this number is possibly very small the data is referred to as ‘sparse’. In the case of sparse data it is common to assume that \( m_i \) are bounded (i.e. \( m_i < M \) for all \( i \in \{1, ..., n\} \)); whereas in dense case is common to assume that \( m_i \) increases with \( n \). The methodology in this thesis is applicable to the case when multiple sets of functional data are observed on either sparse or dense grids of points.

Due to the surge of applications involving data with functional attributes in recent decades, FDA has undergone many theoretical and methodological developments; see Besse and Ramsay (1986), Rice and Silverman (1991), Ferraty et al. (2007), and Horváth and Kokoszka (2012), Mclean et al. (2014) to name a few. For a complete review of FDA methodology we refer the reader to Ramsay and Silverman (2005) and Ferraty et al. (2011).

In Chapter 2 we propose novel methodology for testing that two samples of curves have the same distribution. Then in Chapter 4 we propose novel methodology for function-on-function regression. For each problem, we rely on smoothing the functional data, and employing the main tool in functional data analysis, functional principal components analysis (FPCA). In the following sections we discuss smoothing and FPCA in more detail.

1.2.1 Smoothing Using Basis Function Expansions

In the FDA literature it is common to represent the data using basis function expansions; see Efroymovich (1999) or Ramsay and Silverman (2005) for example. Specifically, let the function \( X(t) \), where we drop the \( i \) index for simplicity, be approximated using a sum of known basis function \( \{B_k(t)\} \) (e.g., B-splines, cubic splines or Fourier basis). That is, \( X(t) \approx \sum_{k=1}^{K} B_k(t) b_k \) where the \( b_k \)'s are the unknown coefficients and \( K \) is large enough to capture the variability of the data. Note that as \( K \) increases also the complexity of the model increases in terms of the number of unknown coefficients needed for estimation using regression techniques. Furthermore, larger values of \( K \) can lead to overfitting of the data. A popular choice for regression is the least square error \( \int \{Y(t) - X(t)\}^2 \, dt \), where again \( Y(t) \) is the response function and \( X(t) \) is
the true underlying function to be estimated.

The pre-specified set of basis functions should be chosen such that the characteristics of the set of functions provides an accurate representation of the features of the data. For example, Fourier basis can be used to capture periodicity on the functions, or splines can be used to characterize localized smooth variations. In Chapter 4, we use B-splines for the representation of a regression function.

B-splines are piecewise polynomial function of degree \( d \) which are non-zero only over a localized region of the domain. Note that the order of each spline is \( d + 1 \) since we also count the constant term. The support of these splines is defined by a set of knots, which are points along \( T \) that split the domain into subintervals. The splines are polynomials of the specified degree over each subinterval with matching derivatives along each knot. The number \( K \) of basis functions, given a set of \( c \) interior knots of degree \( d \), is \( K = d + c + 1 \) (Ramsay and Silverman, 2005). The locations of the knots are often selected to be equally spaced for densely sampled data; however, quantiles of the observed time values can be used as well (Ruppert et al., 2003). Efromovich (1999) provides a comprehensive review of these representations.

1.2.2 Smoothing Using Roughness Penalties

Another method to obtain smooth estimates \( \hat{X}_i(t) \), is to employ roughness penalties which help achieve a balance between how close to the true function \( X_i(t) \) the estimate is (bias) and how wiggly the estimate is (variance). An estimator without any penalty may fit the data very well guaranteeing a low bias, but the estimate may display a lot of local variations hence having a high variance. For the rest of this section, we consider a dense set of grid points and look into estimation of a single curve \( X(t) \), where we drop the \( i \) index for simplicity.

Estimation of \( X(t) \) is often performed via minimization of a loss function which measures closeness to the observed data. A popular choice is the least square error \( \int \{ Y(t) - X(t) \}^2 dt \). Smoothness of the estimator can be controlled by adding a penalty associated with the second derivative \( X''(t) \) of the function. This penalty gives a measure of curvature of \( X(t) \). High values of second derivative correspond to high wiggyness of the function, while a function with \( X''(t) = 0 \) is just a line. The penalized criterion for minimization is given by \( \int \{ Y(t) - X(t) \}^2 dt + \theta \int \{ X''(t) \}^2 dt \), where \( \theta \geq 0 \) is the smoothing parameter. High values of \( \theta \) produce smooth estimates, while smaller value of \( \theta \) produce estimates that resemble the unpenalized results. Traditionally, the smoothing parameter is denoted as \( \lambda \) but we reserve this variable to denote eigenvalues in the next section as well as Chapter 2. In Chapter 4, using such roughness penalties is central to the estimation of a proposed functional regression model, so at that point (with a slight abuse of notation) we will revert to the common notation of using \( \lambda \) to represent the smoothing parameter.
In this context, if we approximate $X(t) \approx \sum_{k=1}^{K} B_k(t) b_k$ where $\{B_k(t)\}$ are basis functions (e.g., B-splines) and $b_k$ are the unknown coefficients, then we can represent the penalty term as, $\int \{X''(t)^2 \} dt = b^\top S b$ where $S$ is a penalty matrix specified by the choice of basis functions such that $S_{kl} = \int B_k''(t) B_l''(t) dt$ (Wood, 2006, 2008). Given that we only observe a set of finite time samples $\{t_j\}$ then the optimization criterion can be expressed as $\sum_j ||Y(t_j) - X(t_j)||^2 + \theta b^\top S b$.

The first term quantifies how well $X(t)$ follows the data, while the second term quantifies smoothness. The regularization parameter $\theta$ provides a balance between the two terms. This representation yields closed form solutions for the minimization criterion for fixed values of $\theta$. Optimal choices of the regularization parameter can be selected using techniques such as cross-validation (Wahba, 1990) or by posing the optimization problem as the fitting of a mixed model in which the smoothing parameter is a variance component, and can be solved using methods such as REML (Wood, 2004, 2006, 2008). This fact will be used in Chapter 4 when estimating functional regression parameters.

### 1.2.3 Functional Principal Components Analysis

For $X(t)$ a square integrable function as in equation (1.1), with continuous mean and covariance function. Mercer’s theorem yields the spectral decomposition of the covariance function, $\Sigma(t, s) = \sum_{k \geq 1} \lambda_k \phi_k(t) \phi_k(s)$ in terms of non-negative eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ and orthogonal eigenfunctions $\phi_k(\cdot)$, with $\int_0^1 \phi_k(t) \phi_{k'}(t) dt = 1(k = k')$, where $1(k = k')$ is the indicator function which is 1 when $k = k'$ and 0 otherwise (Bosq, 2000). We refer to $\{\phi_k\}_k$ as the eigenbasis of $X(t)$ and to $\lambda_k$’s as the corresponding eigenvalues. The decomposition implies that $X(t)$ can be represented via the Karhunen-Loeve (KL) expansion as $X(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_k \phi_k(t)$ where $\xi_k = \int_0^1 \{X(t) - \mu(t)\} \phi_k(t) dt$ are commonly called FPC scores and are uncorrelated random variables with zero mean and variance equal to $\lambda_k$. For practical as well as theoretical reasons (see for example Yao et al. (2005a), Hall et al. (2006), or Di et al. (2009)) the infinite expansion of $X(\cdot)$ is often truncated such that, $X^K(t) = \mu(t) + \sum_{k=1}^{K} \xi_k \phi_k(t)$. It follows that $X^K(t) \rightarrow X(t)$ as $K \equiv K_n \rightarrow \infty$ as $n \rightarrow \infty$, where the convergence is uniform in quadratic mean.

This methodology is used to reduce the dimension of the data in Chapter 2 and to obtain estimates of the smooth trajectories in Chapter 4. In practice, one can consider the observed data as described in model (1.1) are noisy and possibly sampled on an unequal finite grid of points. We generally consider two possible realistic scenarios; 1) the data are observed on a dense grid of points possibly with noise; and 2) the data are observed on a sparse grid of points possibly with noise.

First, consider the situation when the grid of points for each subject is dense and without loss of generality in $[0, 1]$. If the data is sufficiently dense and observed without noise, one could use
pointwise estimators for the mean, $$\hat{\mu}_Y(t) = (1/n) \sum_{i=1}^{n} Y_i(t)$$ and covariance function, $$\hat{\Sigma}_Y(t, s) = (1/n) \sum_{i=1}^{n} \{Y_i(t) - \bar{Y}_i(t)\}\{Y_i(s) - \bar{Y}_i(s)\}$$. If the data are observed with noise as in model (1.1), one can reconstruct the curves $$X_i(t)$$ with negligible error by smoothing the observed functional observations $$\{Y_{ij}\}_{j=1}^{m_i}$$ using local polynomial kernel smoothing (Zhang and Chen, 2007). Let $$\hat{X}_i(\cdot)$$ be the reconstructed trajectories. The main requirement for such reconstruction is that the number of measurements $$m_i$$ tends to infinity at a rate faster than the sample size $$n$$.

Denote by $$\hat{X}_i(t)$$ a generic curve in this sample. Let $$\hat{\mu}(t)$$ be the sample average and let $$\hat{\Sigma}(t, s)$$ be the sample covariance functions of the reconstructed trajectories $$\hat{X}_i(t)$$. Under regularity assumptions these functions are asymptotically identical to the ideal estimators based on the true trajectories (Zhang and Chen, 2007). The spectral decomposition of the estimated covariance function yields the pairs of estimated eigenfunctions and eigenvalues $$\{\hat{\phi}_k(t), \hat{\lambda}_k\}_k$$, with $$\lambda_1 > \lambda_2 > \ldots \geq 0$$. It follows that $$\hat{\xi}_{ik} = \int \{\hat{X}_i(t) - \hat{\mu}(t)\}\hat{\phi}_k(t) dt$$ are consistent estimators of the FPC scores $$\xi_{ik}$$ (Hall et al., 2006; Zhang and Chen, 2007).

For large sample size $$n$$, the distribution of $$\hat{\xi}_{ik}$$ approximates that of $$\xi_{ik}$$; we use this fact to develop a two sample test for functional data in Chapter 2. In applications, $$\hat{\xi}_{ik}$$ can be calculated via numerical integration. The finite truncation $$K$$ of the estimated eigenfunctions $$\{\hat{\phi}_k(t)\}_k$$ can be chosen using model selection based-criteria. There are several methods in the literature to select (or estimate) the finite truncation $$K$$, such as Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) etc.; from our empirical experience the simple criterion based on percentage of explained variance (such as 90% or 95%) gives satisfactory results (Di et al., 2009; Staicu et al., 2010).

Next, consider the situation when the grid of points for each subject is sparse and without loss of generality in $$[0, 1]$$. The sparse setting requires different methodology for several reasons. First, the bounding constraint on the number of repeated observations, $$m_i$$, implies a sparse setting at the curve level and does not provide accurate estimators by smoothing each curve separately. Secondly, estimation of the basis coefficients $$\xi_{ik}$$ via numerical integration is no longer reliable.

Common FPCA-techniques can be applied to reconstruct the underlying subject-trajectories, $$\hat{X}_i(t)$$ from the observed data $$\{Y_{ij} : 1 \leq j \leq m_i\}$$ (Yao et al., 2005a; Di et al., 2009). The key idea is to first obtain estimates of the smooth mean and covariance functions of $$X(t)$$, $$\hat{\mu}(t)$$ and $$\hat{\Sigma}(t, s)$$ respectively. Smooth estimates of the mean of the underlying function $$X(t)$$, can be obtained using the methodology in the previous section or by linear kernel smoothing (Yao et al., 2005a). To obtain estimates of the covariance function one can extend the methods of the previous section to bivariate functions (Wood, 2006) or bivariate kernel smoothing (Yao et al., 2005a).

Once we obtain estimates of the smooth covariance function, the spectral decomposition
yields the eigenfunction/eigenvalue pairs, \{\hat{\phi}_k(\cdot), \hat{\lambda}_k\}_{k \geq 1}$, where $\hat{\lambda}_1 > \hat{\lambda}_2 > \ldots \geq 0$. Next, the variance of the noise is estimated based on the difference between the pointwise variance of the observed data $Y_{ij}$’s and the estimated pointwise variance $\hat{\Sigma}(t, t)$ (Staniswalis and Lee, 1998; Yao et al., 2005a). The finite truncation $K$ of the estimated eigenfunctions $\{\hat{\phi}_k(t)\}_{k}$ can be chosen using model selection based-criteria.

Once the mean function, eigenfunctions, eigenvalues, and noise variance are estimated, the model for the observed data $\{Y_{ij} : 1 \leq j \leq m_i\}$ becomes a linear mixed effects model $Y_{ij} = \hat{\mu}(t_{ij}) + \sum_k \xi_{ik} \hat{\phi}_k(t_{ij}) + \epsilon_{ij}$, where $\text{var}(\xi_{ik}) = \hat{\lambda}_k$ and $\text{var}(\epsilon_{ij}) = \hat{\sigma}_\epsilon^2$. The coefficients $\xi_{ik}$ can be predicted using the conditional expectation formula $\hat{\xi}_{ik} = \hat{E}[\xi_{ik}|(Y_{i1}, \ldots, Y_{im_i})]$. Under the assumption that the responses and errors are jointly Gaussian, the predicted coefficients are in fact the empirical best linear unbiased predictors: $\hat{\xi}_{ik} = \hat{\lambda}_k \hat{\Phi}_i^T (\hat{\Sigma}_i + \hat{\sigma}_\epsilon^2 I_{m_i \times m_i})^{-1} (Y_i - \hat{\mu}_i)$. Here $Y_i$ is the $m_i$-dimensional vector of $Y_{ij}$, $\hat{\mu}_i$ and $\hat{\Phi}_i$ are $m_i$-dimensional vectors with the $j$th entries $\hat{\mu}(t_{ij})$ and $\hat{\phi}(t_{ij})$ respectively, $\hat{\Sigma}_i$ is a $m_i \times m_i$-dimensional matrix with the $(j, j')$th entry equal to $\hat{\Sigma}(t_{ij}, t_{ij'})$, and $I_{m_i \times m_i}$ is the $m_i \times m_i$ identity matrix. Yao et al. (2005a) proved that $\hat{\xi}_{ik}$’s are consistent estimators of $\xi_{ik} = \hat{E}[\xi_{ik}|(Y_{i1}, \ldots, Y_{im_i})]$. In the next chapter we see how this estimator can be used to develop a two-sample test for functional data.
Chapter 2

A Two-Sample Distribution-Free Test for Functional Data with Applications to Diffusion Tensor Imaging

2.1 Background and Motivation

Statistical inference in functional data analysis has been under intense methodological and theoretical development due to the surge of applications involving data with functional features; see Besse and Ramsay (1986), Rice and Silverman (1991), Ramsay and Silverman (2005), Ferraty et al. (2007), and Horváth and Kokoszka (2012) to name a few. Nevertheless, testing the hypothesis that the generating distributions of two sets of curves are identical, when the observed data are noisy and discrete realizations of the curves, has received very little attention. In this chapter, we propose a novel framework based on functional principal component analysis of an appropriate mixture process, which we refer to as marginal functional principal component analysis (marginal FPCA). We present a stable and easy to use computational procedure. Our approach is applicable to a variety of realistic scenarios, such as 1) curves observed at dense or sparse grids of points; with or without measurement error; 2) different sampling designs across the samples; and 3) different sample sizes. The methodology scales well with the total sample size, and it can be extended to test for the equality of more than two samples of curves. Our motivating application is a brain tractography study, where the objective to assess if certain imaging modalities are useful in between patients with multiple sclerosis (MS) and healthy controls with respect to the integrity of signal transmission.
Two-sample hypothesis testing for functional data has been considered in many contexts; ranging from testing for specific types of differences, such as differences in the mean or covariance functions, to testing for overall differences in the cumulative density functions. To detect differences in the mean functions of two independent samples of curves, Ramsay and Silverman (2005) introduced a pointwise t-test, Zhang et al. (2010) presented an $L^2$-norm based test, Horváth et al. (2013) proposed a test based on the sample means of the curves, and Staicu et al. (2014) developed a pseudo-likelihood ratio test. Extensions to $k$ independent samples of curves was discussed in Cuevas et al. (2004), Estévez-Pérez and Vilar (2008), and Laukaitis and Račkauskas (2005), who proposed ANOVA-like testing procedures for testing the equality of mean functions. Recent research also focused on detecting differences in the covariance functions of independent samples of curves: see the factor-based test proposed by Ferraty et al. (2007), the regularized M-test introduced by Kraus and Panaretos (2012), and the chi-squared test proposed by Fremdt et al. (2012).

Literature on testing the equality of the distributions of two samples of functional data observed at discrete grid points and possibly with error is rather scarce and to the best of our knowledge has only been considered previously by Hall and Van Keilegom (2007). The authors proposed a Cramer-von Mises (CVM)-type of test, based on the empirical distributional functionals of the reconstructed smooth trajectories, when functional data are observed on dense designs. Benko et al. (2009) attempted to address this testing problem by first using a functional principal components (FPC) decomposition of the data and then employing a sequential bootstrap test to identify differences in the mean functions, eigenfunctions, and eigenvalues. However, in the proposed form, this test does not account for multiple comparisons and is difficult to study under a variety of scenarios due to its computational expense. Furthermore, even if the multiple comparisons are properly accounted for, the test is still limited to detecting first and second moment changes in the distribution of the FPC scores.

In this chapter, we propose an approach based on the so-called marginal FPCA, which facilitates representation of the curves using the marginal eigenbasis. This reduces the original infinite dimensional two-sample functional testing problem to an approximate simpler finite dimensional testing problem. The methodology is illustrated using the two-sample Anderson-Darling statistic (Pettitt, 1976); however, any other two-sample distribution-free tests can also be used. Our simulation results show that in cases where the approach of Hall and Van Keilegom (2007) applies, our proposed test is considerably more powerful.

Section 2.2 presents the statistical framework and methodology. Then methodology is illustrated under a variety of simulation settings in section 2.4. The simulation results show that in cases where the approach of Hall and Van Keilegom (2007) applies, our proposed test is considerably more powerful. In section 2.5 we apply the methodology to the DTI data of
2.2 Two-Sample Testing for Functional Data

Suppose we observe data that arise from two groups, \(Y_{1ij}, t_{1ij} : i \in \{1, \ldots, n_1\} \) and \(j \in \{1, \ldots, m_{1i}\}\) and \(Y_{2ij}, t_{2ij} : i \in \{1, \ldots, n_2\} \) and \(j \in \{1, \ldots, m_{2i}\}\), where \(t_{1ij}, t_{2ij} \in T\) for some bounded and closed interval \(T\); for simplicity take \(T = [0, 1]\). The notation of the time-points, \(t_{1ij}\) and \(t_{2ij}\), allows for different observation points in the two groups. It is assumed that the \(Y_{1ij}\)'s and \(Y_{2ij}\)'s are independent realizations of two underlying (stochastic) processes observed with noise on a finite grid of points. Specifically, consider the model,

\[
Y_{1ij} = X_{1i}(t_{1ij}) + \epsilon_{1ij}, \quad \text{and} \quad Y_{2ij} = X_{2i}(t_{2ij}) + \epsilon_{2ij},
\]

where \(X_{1i}(\cdot) \overset{iid}{\sim} X_1(\cdot)\) and \(X_{2i}(\cdot) \overset{iid}{\sim} X_2(\cdot)\) are independent and square-integrable random functions over \(T\), for some underlying (latent) random processes \(X_1(\cdot)\) and \(X_2(\cdot)\). It is assumed that \(X_1(\cdot)\) and \(X_2(\cdot)\) have unknown continuous mean and continuous and positive semi-definite covariance functions. The measurement errors, \(\{\epsilon_{1ij}\}_{i,j}\) and \(\{\epsilon_{2ij}\}_{i,j}\), are independent and identically distributed with mean zero, and with variances \(\sigma^2_{\epsilon_1}\) and \(\sigma^2_{\epsilon_2}\) respectively, and are independent of \(X_{1i}(\cdot)\) and \(X_{2i}(\cdot)\). Our objective is to test the null hypothesis,

\[
H_0 : X_1(\cdot) \overset{d}{=} X_2(\cdot)
\]

versus the alternative hypothesis \(H_A : X_1(\cdot) \overset{d}{\neq} X_2(\cdot)\), where \(\overset{d}{=}\) denotes the processes on either side have the same distribution. Our interest is to develop non-parametric and computationally inexpensive methods to test (2.2).

Since \(X_1(\cdot)\) and \(X_2(\cdot)\) are processes defined over a continuum, testing (2.2) implies testing the null hypothesis that two infinite dimensional objects have the same generating distribution. This is different from two-sample testing in a multivariate framework, where the dimension of the random objects of interest is finite and the same. In the case where the sampling design is common to all the subjects (i.e. \(t_{1ij} = t_{2ij} = t_j\) and \(m_{1i} = m_{2i} = m\)), the dimension of the testing problem could potentially be reduced by testing an approximate null hypothesis - that the multivariate distribution of the processes evaluated at the observed grid points are equal. Multivariate testing procedures (eg. Aslan and Zech (2005); Friedman and Rafsky (1979); Read and Cressie (1988)) could be employed in this situation. However, these procedures have only been illustrated for cases when \(m = 4\) or \(5\) in our notation. In dense functional data, the number of unique time-points, \(m\), is orders of magnitude larger, often even larger than the sample size.

Recent research has approached this problem using functional data analysis based tech-
niques. For example Hall and Van Keilegom (2007) propose an extension of the Cramer-von Mises (CVM) test from multivariate statistics, and use bootstrap to approximate the null distribution of the test. Benko et al. (2009) consider a common functional principal components model for the two samples and proposed bootstrap procedures to test for the equality of the corresponding model components. Both approaches rely on bootstrap techniques, making it often numerically unfeasible to perform extensive empirical power analysis for large sample sizes.

In this chapter we consider testing hypothesis (2.2) by employing basis function expansions, using the eigenbasis of the marginal covariance operator of the two latent processes, \(X_1(\cdot)\) and \(X_2(\cdot)\). This facilitates dimension reduction of the functional objects and simplification of the hypothesis testing (2.2) to two-sample multivariate testing of the equality of the distributions of the basis coefficients. Furthermore, it reduces the testing (2.2) to a sequence of two-sample tests for the equality of univariate distributions combined with a multiple testing correction. The proposed procedure is computationally inexpensive and scales well with large sample sizes.

To begin with, we describe how to test hypothesis (2.2) under the assumption that the curves are observed entirely and without noise (Hall et al., 2006). Extension to practical settings is discussed in Section 2.3. Consider two sets of independent curves \(\{X_{11}(\cdot), \ldots, X_{n_1}(\cdot)\}\) and \(\{X_{21}(\cdot), \ldots, X_{2n_2}(\cdot)\}\), defined on \([0,1]\). Assume \(X_{11}(\cdot) \sim X_1(\cdot)\) and \(X_{21}(\cdot) \sim X_2(\cdot)\) are square integrable and have continuous mean and covariance functions respectively.

Our methodology is developed under the assumption that both \(n_1, n_2 \to \infty\) such that \(\lim_{n_1,n_2 \to \infty} n_1/(n_1 + n_2) = p \in (0,1)\). Let \(X(\cdot)\) be the mixture process of \(X_1(\cdot)\) and \(X_2(\cdot)\) with mixture probabilities \(p\) and \(1-p\) respectively. Furthermore, let \(Z\) be a binary random variable taking values in \(\{1,2\}\) such that \(P(Z = 1) = p\). Then \(X_1(\cdot)\) is the conditional process \(X(\cdot)\) given \(Z = 1\), and \(X_2(\cdot)\) is the conditional process \(X(\cdot)\) given \(Z = 2\). It follows that \(X(\cdot)\) is square integrable, and its marginal distribution has continuous mean and positive semi-definite covariance functions. Let \(\mu(t) = E[X(t)]\) be the (marginal) mean function and let \(\Sigma(t,s) = \text{cov}\{X(t), X(s)\}\) be the (marginal) covariance function of \(X(\cdot)\). Mercer’s theorem yields the spectral decomposition of the marginal covariance function, \(\Sigma(t,s) = \sum_{k\geq 1} \lambda_k \phi_k(t)\phi_k(s)\) in terms of non-negative eigenvalues \(\lambda_1 \geq \lambda_2 \geq \ldots \geq 0\) and orthogonal eigenfunctions \(\phi_k(\cdot)\), with \(\int_0^1 \phi_k(t)\phi_{k'}(t)dt = 1(k = k')\), where \(1(k = k')\) is the indicator function which is 1 when \(k = k'\) and 0 otherwise (Bosq, 2000). We refer to \(\{\phi_k\}_k\) as the marginal eigenbasis of \(X_1(\cdot)\) and \(X_2(\cdot)\), and to \(\lambda_k\)’s as the corresponding marginal eigenvalues. The decomposition implies that \(X(\cdot)\) can be represented via the KL expansion as \(X(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_k \phi_k(t)\) where \(\xi_k = \int_0^1 \{X(t) - \mu(t)\}\phi_k(t)dt\) are commonly called FPC scores and are uncorrelated random variables with zero mean and variance equal to \(\lambda_k\). For practical as well as theoretical reasons (see for example Yao et al. (2005a), Hall et al. (2006), or Di et al. (2009)) the infinite expansion of \(X(\cdot)\) is often truncated. Let \(X^K(t) = \mu(t) + \sum_{k=1}^{K} \xi_k \phi_k(t)\) be the truncated KL expansion of \(X(\cdot)\).
It follows that $X^K(t) \to X(t)$ as $K \equiv K_n \to \infty$ as $n \to \infty$, where the convergence is uniform in quadratic mean which leads to the lemma below. Define $X^K_z(t) = \mu(t) + \sum_{k=1}^{K} \xi_{zk} \phi_k(t)$ to be the finite-dimensional approximation of $X_z(t)$ such that $\xi_{zk} = \int_0^t (X_z(t) - \mu(t)) \phi_k(t) dt$ for $k \geq 1$ and $z = 1, 2$.

**Lemma:** If $\sup_t E[(X(t) - X^K(t))^2] \to 0$ as $K \to \infty$ then $E[(X_z(t) - X^K_z(t))^2] \to 0$ as $K \to \infty$, uniformly in $t$, for $z \in \{1, 2\}$

This result easily follows from noting that the mixture process $X(t)$ is defined conditionally on the variable $Z$ so that, $\sup_t E(E[(X(t) - X^K(t))^2|Z]) = p \sup_t E[(X_1(t) - X^K_1(t))^2] + (1 - p) \sup_t E[(X_2(t) - X^K_2(t))^2] \geq 0$. Notice that in general, $\xi_{zk}$’s are not necessarily uncorrelated or have mean zeros. The lemma states that $E[(X_z(t) - \mu(t) - \sum_{k=1}^{K} \xi_{zk} \phi_k(t))^2] \to 0$ for $K \to \infty$, uniformly in $t$, and for $z \in \{1, 2\}$. Furthermore, if null hypothesis (2.2), holds true then it follows that $X^K_1(\cdot) \overset{d}{=} X^K_2(\cdot)$, for each finite truncation $K$. Let $K$ be a suitable finite-dimensional truncation such that $X^K(\cdot)$ approximates $X(\cdot)$ accurately using $L^2$ norm; it follows that $X^K_z(\cdot)$ approximates well $X_z(\cdot)$ for $z = 1, 2$. Since $\{\phi_1(\cdot), \phi_2(\cdot), \ldots\}$ is the eigenbasis of the covariance of the marginal distribution of $X(\cdot)$ we refer to the analysis based on this basis function expansion by ‘marginal FPCA’.

**Proposition:** For large enough $K$, the null hypothesis (2.2) can be reduced to

$$H^K_0: \{\xi_{1k}\}_{k=1}^{K} \overset{d}{=} \{\xi_{2k}\}_{k=1}^{K}. \quad (2.3)$$

Here, the superscript $K$ in $H^K_0$ is to emphasize the dependence of the reduced null hypothesis on the finite truncation $K$. The proof of this result is based on the observation that for finite truncation $K$, we have $X^K_1(\cdot) \overset{d}{=} X^K_2(\cdot)$ if only if the multivariate distributions of basis coefficients, $\{\xi_{1k}\}_{k=1}^{K}$ and $\{\xi_{2k}\}_{k=1}^{K}$, are the same.

One possible approach to test hypothesis (2.3) is to consider two-sample multivariate procedures; see for example Wei and Lachin (1984), Schilling (1986) or Bohm and Zech (2010), Ch.10. For simplicity, we consider multiple two-sample univariate tests combined with a multiple comparison adjustment (e.g. Bonferroni correction). In particular, testing the null hypothesis (2.3) can be carried by multiple testing of the null hypotheses $H^K_{0k}$, for $k = 1, \ldots, K$, where

$$H^K_{0k}: \xi_{1k} \overset{d}{=} \xi_{2k}. \quad (2.4)$$

There are several common univariate two-sample tests; for example the Kolmogorov-Smirnov test (KS, Massey Jr (1951)) or the Anderson-Darling test (AD, Pettitt (1976)). KS and AD tests are both capable of detecting higher order moment shifts between the two univariate distributions, by using differences in the empirical cumulative distributions. Empirical studies have shown that AD test tends to have higher power than KS test (Stephens, 1974; Bohm and
we apply it to functional data we call this test the Functional Anderson-Darling (FAD) applications, as the true smooth trajectories $X$ are known. Then, the corresponding basis coefficients $\xi_{zik}$’s can be determined as $\xi_{zik} = \int \{ X_{zik}(t) - \mu(t) \} \phi_k(t) dt$ and $\xi_{2ik} = \int \{ X_{2i}(t) - \mu(t) \} \phi_k(t) dt$. Let $\tilde{F}_1(\cdot)$ and $\tilde{F}_2(\cdot)$ be the corresponding empirical conditional distribution functions of the $\xi_{zik}$’s and $\xi_{2zik}$’s respectively. The AD test statistic is defined as,

$$AD_k^2 = \frac{n_1n_2}{n} \int_{-\infty}^{\infty} \{ \frac{\tilde{F}_1(x) - \tilde{F}_2(x)}{\tilde{F}_k(x)\{1 - \tilde{F}_k(x)\}} \} d\tilde{F}_k(x),$$

where $n = n_1 + n_2$ and $\tilde{F}_k(x) = \{n_1\tilde{F}_{1k}(x) + n_2\tilde{F}_{2k}(x)\}/n$ (Pettitt, 1976; Scholz and Stephens, 1987). Under the null hypothesis $H_k^{K}$ of (2.4), the AD test statistic, $AD_k^2$, converges to the same limiting distribution as the AD for one sample (Pettitt, 1976). Given a univariate two-sample test, define an $\alpha$-level testing procedure to test hypothesis (2.2) as follows: hypothesis (2.2) is rejected if $\min_{1 \leq k \leq K} p_k \leq (\alpha/K)$, where $p_k$ is the $p$-value which is obtained using the chosen univariate two-sample test for $H_{k0}$, for $k = 1, \ldots, K$. The use of the Bonferroni correction ensures that the testing procedure maintains its nominal size, conditional on the truncation level $K$. Because we apply it to functional data we call this test the Functional Anderson-Darling (FAD). The proposed testing methodology allows us to extend any univariate testing to the case of functional data. Of course, any advantages or drawbacks of the univariate tests, such as the ability to detect higher order moment shifts or weak power in small sample sizes, will carry over to the functional extension.

### 2.3 Extension to Practical Situations

Extension of the testing procedure described in Section 2.2 is not straightforward in practical applications, as the true smooth trajectories $X_i(\cdot)$ and moreover the true scores $\xi_{ik}$ are not directly observable. Instead the observed data as described in model (2.1) are noisy and possibly sampled on an unequal finite grid of points. Let $Z_i$ be the variable that denotes the group membership of the $i$th curve. To address this challenge, we propose to replace $\xi_{zik}$ from the previous section by appropriate estimators, $\hat{\xi}_{zik}$, and thus test the hypotheses $H_{0k}$ in (2.4) using $\hat{\xi}_{zik}$’s instead of $\xi_{zik}$’s for $z = 1, 2$. Define by $\hat{AD}_k^2$ the statistic analogous to $AD_k^2$, with the basis coefficients $\hat{\xi}_{1ik}$ and $\hat{\xi}_{2ik}$ replacing $\xi_{1ik}$ and $\xi_{2ik}$, respectively. Intuitively, we expect that if the null hypothesis $H_{k0}$ (2.4) is true, then the asymptotic distribution of $\hat{AD}_k^2$ is the same as the asymptotic null distribution of $AD_k^2$.

Intuitively, our logic is based on the result that under null hypothesis (2.2) $\hat{\xi}_{1ik} - \hat{\xi}_{2ik} \xrightarrow{p} 0$ as $n \to \infty$ where $\xrightarrow{p}$ denotes convergence in probability, for $k = 1, \ldots, K$. Thus, to test (2.4) one can use the proposed testing procedure described in the previous section, but with the
and eigenvalues estimated marginal covariance function yields the pairs of estimated marginal eigenfunctions based on the true trajectories (Zhang and Chen, 2007). The spectral decomposition of the Under regularity assumptions these functions are asymptotically identical to the ideal estimators \( \xi \) are consistent estimators of the FPC scores \( \hat{\xi} \) is 

First, consider the situation when the grid of points for each subject is dense in \([0, 1]\), that is \( m_{1i} \) and \( m_{2i} \) are very large. Zhang and Chen (2007) proved that one can reconstruct the curves \( X_i(t) \) with negligible error by smoothing the observed functional observations \( \{Y_{ij}\}_{j=1}^{m_{1i}} \) using local polynomial kernel smoothing. Let \( \hat{X}_{1i}(\cdot) \) and \( \hat{X}_{2i}(\cdot) \) be the reconstructed trajectories in group one and two respectively. The main requirement for such reconstruction is that the number of measurements \( m_{1i} \) and \( m_{2i} \) for all subjects tends to infinity at a rate faster than the sample sizes \( n_1 \), and \( n_2 \) respectively.

Consider the pooled sample \( \{\hat{X}_{1i}(\cdot) : i = 1, \ldots, n_1\} \cup \{\hat{X}_{2i}(\cdot) : i = 1, \ldots, n_2\} \) and denote by \( \hat{X}_i(t) \) a generic curve in this sample. Let \( \hat{\mu}(t) \) be the (marginal) sample average and let \( \hat{\Sigma}(t, s) \) be the (marginal) sample covariance functions of the reconstructed trajectories \( \hat{X}_i(t) \). Under regularity assumptions these functions are asymptotically identical to the ideal estimators based on the true trajectories (Zhang and Chen, 2007). The spectral decomposition of the estimated marginal covariance function yields the pairs of estimated marginal eigenfunctions and eigenvalues \( \{\hat{\phi}_k(t), \hat{\lambda}_k\}_k \), with \( \lambda_1 > \lambda_2 > \ldots \geq 0 \). It follows that \( \hat{\xi}_{ik} = \int \{\hat{X}_i(t) - \hat{\mu}(t)\} \hat{\phi}_k(t) dt \) are consistent estimators of the FPC scores \( \xi_{ik} \) (Hall et al., 2006; Zhang and Chen, 2007); \( \hat{\xi}_{1ik} = \hat{\xi}_{ik} \) if \( Z_i = 1 \) and \( \hat{\xi}_{2ik} = \hat{\xi}_{ik} \) if \( Z_i = 2 \). Therefore, for large sample sizes \( n_1 \) and \( n_2 \), the distribution of \( \hat{\xi}_{zik} \) approximates that of \( \xi_{zik} \). In applications, \( \hat{\xi}_{ik} \) can be calculated via numerical integration. Therefore, \( \hat{\xi}_{ik} \) are used for testing hypothesis (2.3). The finite truncation \( K \) of the estimated eigenfunctions \( \{\hat{\phi}_k(t)\}_k \) can be chosen using model selection based-criteria. We found that the cumulative explained variance criterion (Di et al., 2009; Staicu et al., 2010) works very well, for estimation of \( K \), in practice.

Next consider the situation when the grid of points for each subject is sparse in \([0, 1]\), so that, \( m_{1i}, m_{2i} \) are possibly very small. The sparse setting requires different methodology for several reasons. First, the bounding constraint on the number of repeated observations, \( m_{1i} \), and \( m_{2i} \) respectively, implies a sparse setting at the curve level and does not provide accurate estimators by smoothing each curve separately. Secondly, estimation of the basis coefficients \( \xi_{ik} \) via numerical integration is no longer reliable. Instead, we consider the pooled sample \( \{Y_{1ij} : i, j\} \cup \{Y_{2ij} : i, j\} \), and let \( \{Y_{ij} : j \in \{1, \ldots, m_i\}\} \) be a generic observed profile in this set. The observed measurements \( \{Y_{ij} : j \in \{1, \ldots, m_i\}\}_i \) are viewed as independent and identically distributed realizations of the an appropriate mixture stochastic process, that are observed at finite grids \( \{t_{i1}, \ldots, t_{im_i}\} \) and contaminated with error. Specifically it is implied that \( Y_{ij} = \)
\[ X_i(t_{ij}) + \epsilon_{ij}, \] for \( X_i(\cdot) \) which is the mixture process as described earlier with \( E[X(t)] = \mu(t) \) and \( \text{cov}\{X(t), X(s)\} = \Sigma(t, s) \). Here \( \epsilon_{ij} \)'s are independent and identically distributed measurement error with zero mean and variance \( \sigma^2 \).

Common FPCA-techniques can be applied to reconstruct the underlying subject-trajectories, \( \hat{X}_i(\cdot) \) from the observed data \( \{Y_{ij} : 1 \leq j \leq m_i\} \) (Yao et al., 2005a; Di et al., 2009). The key idea is to first obtain estimates of the marginal smooth mean and covariance functions, \( \hat{\mu}(t) \) and \( \hat{\Sigma}(t, s) \) respectively. The spectral decomposition of the estimated marginal covariance yields the marginal eigenfunction/eigenvalue pairs, \( \{\hat{\phi}_k(\cdot), \hat{\lambda}_k\}_{k \geq 1} \), where \( \hat{\lambda}_1 > \hat{\lambda}_2 > \ldots \geq 0 \). Next, the variance of the noise is estimated based on the difference between the pointwise variance of the observed data \( Y_{ij} \)'s and the estimated pointwise variance \( \hat{\Sigma}(t, t) \) (Staniswalis and Lee, 1998; Yao et al., 2005a). There are several methods in the literature to select (or estimate) the finite truncation \( K \), such as Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) etc.; from our empirical experience the simple criterion based on percentage of explained variance (such as 90% or 95%) gives satisfactory results. In our simulation experiments we use the cumulative explained variance to select \( K \). Sensitivity with regards to this parameter is studied in Section 2.4.

Once the marginal mean function, marginal eigenfunctions, eigenvalues, and noise variance are estimated, the model for the observed data \( \{Y_{ij} : 1 \leq j \leq m_i\} \) becomes a linear mixed effects model \( Y_{ij} = \hat{\mu}(t_{ij}) + \sum_k \xi_{ik} \hat{\phi}_k(t_{ij}) + \epsilon_{ij}, \) where \( \text{var}(\xi_{ik}) = \hat{\lambda}_k \) and \( \text{var}(\epsilon_{ij}) = \hat{\sigma}^2 \). The coefficients \( \xi_{ik} \) can be predicted using the conditional expectation formula \( \hat{\xi}_{ik} = \hat{\hat{E}}[\xi_{ik}|(Y_{i1}, \ldots, Y_{im})] \). Under the assumption that the responses and errors are jointly Gaussian, the predicted coefficients are in fact the empirical best linear unbiased predictors: \( \hat{\xi}_{ik} = \hat{\lambda}_k \hat{\Phi}_i^T (\hat{\Sigma}_i + \hat{\sigma}^2 I_{m_i \times m_i})^{-1} (Y_i - \hat{\hat{\mu}}_i) \).

Here \( Y_i \) is the \( m_i \)-dimensional vector of \( Y_{ij} \); \( \hat{\mu}_i \) and \( \hat{\Phi}_i \) are \( m_i \)-dimensional vectors with the \( j \)th entries \( \hat{\mu}(t_{ij}) \) and \( \hat{\phi}(t_{ij}) \) respectively, \( \hat{\Sigma}_i \) is a \( m_i \times m_i \)-dimensional matrix with the \( (j, j') \)th entry equal to \( \hat{\Sigma}(t_{ij}, t_{ij'}) \), and \( I_{m_i \times m_i} \) is the \( m_i \times m_i \) identity matrix. Yao et al. (2005a) proved that \( \hat{\xi}_{ik}'s \) are consistent estimators of \( \hat{\xi}_{ik} = E[\xi_{ik}|(Y_{i1}, \ldots, Y_{im})] \). Define \( \hat{\xi}_{zik} = \hat{\xi}_{ik} \) to be the conditional expectation - based coefficients when \( Z_i = z \) and similarly define \( \hat{\xi}_{zik} = \hat{\xi}_{ik} \) if \( Z_i = z \), their estimated counterparts. If \( \xi_{1i} \overset{d}{=} \xi_{2i} \) it follows that \( \hat{\xi}_{1i} \overset{d}{=} \hat{\xi}_{2i} \); the reverse implication is true under a joint Gaussian assumption. It follows that for large sample sizes \( n_1 \) and \( n_2 \) the sampling distribution of \( \hat{\xi}_{1ik}'s \) and \( \hat{\xi}_{2ik}'s \) are the same as those of \( \xi_{1ik}'s \) and \( \xi_{2ik}'s \), respectively. Therefore, we can use \( \hat{\xi}_{zik}'s \) to test hypothesis (2.4).

### 2.4 Simulation Studies

The performance of the proposed testing procedure is presented under a variety of settings and for varying sample sizes. We focus on a small number of FPCs first and later study more complex data settings. Section 2.4.1 studies Type I error rate of the FAD test and the sensitivity to sample
size and the percentage of explained variance, \( \tau \), used to estimate the parameter, \( K \). Section 2.4.2 provides a numerical comparison of the proposed approach with the closest available competitor - the Cramér-von Mises (CVM) -type test introduced by Hall and Van Keilegom (2007).

### 2.4.1 Type One Error and Power Performance

We construct datasets \( \{(t_{1ij}, Y_{1ij}) : j \} \) and \( \{(t_{2ij}, Y_{2ij}) : j \} \) using model (2.1) for \( t_{1ij} = t_{2ij} = t_j \) observed on an equally spaced grid of \( m = 100 \) points in \([0, 1]\). Here \( X_{1i}(t) = \mu_1(t) + \sum_k \phi_{1k}(t)\xi_{1ik} \) and \( X_{2i}(t) = \mu_2(t) + \sum_k \phi_{2k}(t)\xi_{2ik} \), where \( \phi_{11}(t) = \phi_{21}(t) = \sqrt{2} \sin(2\pi t) \), \( \phi_{12}(t) = \phi_{22}(t) = \sqrt{2} \cos(2\pi t) \) and so on, are the Fourier basis functions, \( \xi_{1ik} \) and \( \xi_{2ik} \) are uncorrelated respectively with var\(\xi_{1ik} = \lambda_{1k} \), var\(\xi_{2ik} = \lambda_{2k} \) and \( \lambda_{1k} = \lambda_{2k} = 0 \) for \( k \geq 4 \). We set \( \epsilon_{1ij} \sim N(0, 0.25) \) and \( \epsilon_{2ij} \sim N(0, 0.25) \). The setting \( \phi_{2k}(t) = \phi_{k}(t) \) allows us to study different types of departures from the null hypothesis of equal distribution of the underlying processes (see settings A, B, C). In Section 4.2 we consider generating processes \( X_{1i}() \) and \( X_{2i}() \) that have different basis functions representations.

The FAD test is employed to test the null hypothesis (2.2); the marginal mean functions, the marginal eigenbasis functions, and furthermore the corresponding basis coefficients are estimated using the methods described in Section 2.3. The number of marginal eigenbasis functions is estimated using the percentage of explained variance, \( \tau \), for the pooled data. We use \( \tau = 95\% \) in our simulation experiments; as we illustrate next the FAD test is robust to this parameter. The estimates for all the model components, including the basis coefficients, are obtained using the \( R \) package \texttt{refund} (Crainiceanu et al., 2012). Next, the \( R \) package \texttt{AD} (Scholz, 2011) is used to test the equality of the corresponding univariate distributions for each pair of basis coefficients. The Bonferroni multiple testing correction is used to maintain the desired level of the FAD test. The null hypothesis is rejected/not rejected according to the approach described in Section 2.2. All the results in this section are based on \( \alpha = 0.05 \) significance level.

First, we assess the Type I error rate for various threshold parameter values, \( \tau \). For simplicity set \( \mu_1(t) = \mu_2(t) = 0 \) for all \( t \) and consider \( \xi_{1ik}, \xi_{2ik} \sim N(0, \lambda_k) \), for \( \lambda_1 = 10, \lambda_2 = 5, \) and \( \lambda_3 = 2 \) and \( \lambda_k = 0, \forall k > 3 \). Type I error rate is studied for varying \( \tau \) from 80\% to 99\% and for increasing equal/unequal sample sizes. For the simulated data, the proposed criterion to select the number of marginal eigenfunctions yields \( K = 2 \) if \( \tau = 80\% \) and \( K = 3 \) when \( \tau = 99\% \). The table analyzes the sensitivity to \( \tau \); different values of \( \tau \) may result in different choices for \( K \), which in turn may lead to different type I error estimates. Table 2.1 displays the empirical size of the FAD test using varying thresholds \( \tau \) for the case when the sample sizes are equal or unequal, with an overall size increasing from 200 to 2000. The results are based on 5000 MC replications. They show that, as expected, the size of the test is not sensitive to the threshold
\( \tau \); the size is close to the nominal level \( \alpha = 0.05 \) in all cases. Additionally, we investigated numerically the effect of the threshold on the power of the tests, and found that it has very little effect on the power (results not reported).

Next, we study the power performance of the FAD test with \( \tau = 0.95 \). The distribution of the true processes under the alternative is described by the mean functions, as well as by the distributions of the basis coefficients. The following scenarios allow us to study the FAD test for specific types of changes in the two data sets. Setting A corresponds to deviations in the mean functions, settings B and C correspond to deviations in the second moment and third moment, respectively of the corresponding distribution of the first set of basis coefficients. Throughout this section it is assumed that \( \lambda_{1k} = \lambda_{2k} = 0 \) for all \( k \geq 3 \).

[A] **Mean Shift:** Set the mean functions as \( \mu_1(t) = t \) and \( \mu_2(t) = t + \delta t^3 \). Generate the coefficients as \( \xi_{1i1}, \xi_{2i1} \sim N(0, 10), \xi_{1i2}, \xi_{2i2} \sim N(0, 5) \). The index \( \delta \) controls the departure in the mean behavior of the two distributions.

[B] **Variance Shift:** Set \( \mu_1(t) = \mu_2(t) = 0 \). Generate the coefficients \( \xi_{1i1} \sim N(0, 10), \xi_{2i1} \sim N(0, 10 + \delta), \) and \( \xi_{1i2}, \xi_{2i2} \sim N(0, 5) \). Here \( \delta \) controls the difference in the variance of the first basis coefficient between the two sets.

[C] **Skewness Shift:** \( \xi_{1i1} \sim T_4(0, 10) \) and \( \xi_{2i1} \sim ST_4(0, 10, 1 + \delta) \), and \( \xi_{1i2}, \xi_{2i2} \sim T_4(0, 5) \). Here, \( T_4(\mu, \sigma) \) denotes the common students T distribution with 4 degrees of freedom, that is standardized to have mean \( \mu \) and standard deviation \( \sigma \) and \( ST_4(\mu, \sigma, \gamma) \) is the standardized skewed T distribution (Wurtz et al., 2006) with 4 degrees of freedom, mean \( \mu \), standard deviation \( \sigma \), and shape parameter \( 0 < \gamma < \infty \) which models skewness. The shape parameter \( \gamma \) is directly related to the skewness of this distribution and the choice \( \gamma = 1 \) for corresponds to the symmetric T distribution. Thus index \( \delta \) controls the difference in the skewness of distribution of the first basis coefficient. In section 2.4.2 empirical power is studied for a broader class of distribution shifts.

For all the settings, \( \delta = 0 \) corresponds to the null hypothesis, that the two samples of curves have the same generating distribution, whereas \( \delta > 0 \) corresponds to the alternative hypothesis, that the two sets of curves have different distributions. Thus \( \delta \) indexes the departure from the null hypothesis, and it will be used to depict empirical power curves. The estimated power is based on 1000 MC replications. Results are presented in Figure 2.1 for the case of equal/unequal sample sizes in the two groups, and for various total sample sizes.

Column A of Figure 2.1 displays the empirical power curves of the FAD test when the mean discrepancy index \( \delta \) ranges from 0 to 8. It appears that the performance of the power is affected more by the combined sample size, \( n = n_1 + n_2 \), than the magnitude of each sample size \( n_1 \) or \( n_2 \). Column B shows the empirical power, when the variance discrepancy index \( \delta \) ranges from 0 to 70. The empirical power increases at a faster rate for equal sample sizes than unequal sample sizes, when the total sample size is the same. However, the differences become less pronounced as the total sample size increases. Finally, column C displays the power behavior for observed
Figure 2.1: Empirical power curves under simulation setting A (leftmost panels), B (middle panels) and C (rightmost panels). The top panels consist of the case in which both sets of curves have equal sample sizes and the bottom panels correspond to unequal sample sizes as displayed in the legends. The overall sample size $n = n_1 + n_2$ varies from $n = 200$ to $n = 1000$. The maximum standard error is 0.007.

Table 2.1: Estimated Type I Error rate of FAD Test, based on 5000 replications for several threshold values, $\tau$, leading to different estimates of the truncation parameter $K$.

<table>
<thead>
<tr>
<th>$(n_1, n_2)$</th>
<th>$\tau$</th>
<th>0.80</th>
<th>0.85</th>
<th>0.90</th>
<th>0.95</th>
<th>0.99</th>
</tr>
</thead>
<tbody>
<tr>
<td>(100,100)</td>
<td></td>
<td>0.056</td>
<td>0.056</td>
<td>0.059</td>
<td>0.060</td>
<td>0.060</td>
</tr>
<tr>
<td>(200,200)</td>
<td></td>
<td>0.050</td>
<td>0.050</td>
<td>0.052</td>
<td>0.053</td>
<td>0.053</td>
</tr>
<tr>
<td>(300,300)</td>
<td></td>
<td>0.053</td>
<td>0.053</td>
<td>0.049</td>
<td>0.049</td>
<td>0.049</td>
</tr>
<tr>
<td>(500,500)</td>
<td></td>
<td>0.049</td>
<td>0.049</td>
<td>0.050</td>
<td>0.050</td>
<td>0.050</td>
</tr>
<tr>
<td>(1000,1000)</td>
<td></td>
<td>0.055</td>
<td>0.055</td>
<td>0.058</td>
<td>0.058</td>
<td>0.058</td>
</tr>
<tr>
<td>(50,150)</td>
<td></td>
<td>0.055</td>
<td>0.055</td>
<td>0.058</td>
<td>0.058</td>
<td>0.057</td>
</tr>
<tr>
<td>(100,300)</td>
<td></td>
<td>0.053</td>
<td>0.053</td>
<td>0.049</td>
<td>0.049</td>
<td>0.049</td>
</tr>
<tr>
<td>(150,450)</td>
<td></td>
<td>0.048</td>
<td>0.048</td>
<td>0.055</td>
<td>0.054</td>
<td>0.054</td>
</tr>
<tr>
<td>(250,750)</td>
<td></td>
<td>0.051</td>
<td>0.051</td>
<td>0.054</td>
<td>0.054</td>
<td>0.054</td>
</tr>
<tr>
<td>(500,1500)</td>
<td></td>
<td>0.055</td>
<td>0.055</td>
<td>0.053</td>
<td>0.053</td>
<td>0.053</td>
</tr>
</tbody>
</table>
data generated under setting C for \( \delta \) between 0 and 6. The \texttt{rstd} and \texttt{rsstd} functions in the \textit{R} package \texttt{fgarch} (Wurtz et al., 2006) are used to generate random data from a standardized \( T \) and standardized skewed \( T \) distribution respectively. For moderate sample sizes, irrespective of their equality, the probability of rejection does not converge to 1 no matter how large \( \delta \) is; see the results corresponding to a total sample size equal to \( n = 200 \) or 400. This is in agreement with our intuition that detecting differences in higher order moments of the distribution becomes more difficult and requires increased sample sizes. In contrast, for larger total sample sizes, the empirical power curve has a fast rate of increase.

### 2.4.2 Comparison with an Alternative Approach

To the authors’ best knowledge Hall and Van Keilegom (2007) is the only available alternative method that considers hypothesis testing that the distributions of two samples of curves are the same, when the observed data are noisy and discrete realizations of the latent curves. Their methods are presented for dense sampling designs only; thus we restrict the comparison to this design only. In this section we compare the performance of their proposed Cramer-von Mises (CVM) - type test, based on the empirical distribution functions after local-polynomial smoothing of the two samples of curves, with our FAD test.

We generate data \( \{(t_{1ij}, Y_{1ij}) : j\}^{n_1}_{i=1} \) and \( \{(t_{2ij}, Y_{2ij}) : j\}^{n_2}_{i=1} \) exactly as in Hall and Van Keilegom (2007), and for completeness we describe it next: \( Y_{1ij} = X_{1i}(t_{1ij}) + \epsilon_{1ij} \) and \( Y_{2ij} = X_{2i}(t_{2ij}) + \epsilon_{2ij} \), where \( \epsilon_{1ij} \sim iid N(0,0.01) \), \( \epsilon_{2ij} \sim iid N(0,0.09) \), \( X_{1i}(t) = \sum_{k=1}^{15} e^{-k/2}N_{k1i}\psi_k(t) \) and \( X_{2i}(t) = \sum_{k=1}^{15} e^{-k/2}N_{k2i}\psi_k(t) + \delta \sum_{k=1}^{15} k^{-2}N_{k22i}\psi_k^*(t) \), such that \( N_{k1i}, N_{k2i}, N_{k22i} \sim iid N(0,1) \). Here \( \psi_1(t) \equiv 1 \) and \( \psi_k(t) = \sqrt{2}\sin\{k\pi t\} \) are orthonormal basis functions. Also \( \psi_k^*(t) \equiv 1 \), \( \psi_k^*(t) = \sqrt{2}\sin\{(k-1)\pi(2t-1)\} \) if \( k > 1 \) is odd and \( \psi_k^*(t) = \sqrt{2}\cos\{(k-1)\pi(2t-1)\} \) if \( k > 1 \) is even. As before the index \( \delta \) controls the deviation from the null hypothesis; \( \delta = 0 \) corresponds to the null hypothesis, that the two samples have identical distribution. Finally, the sampling design for the curves is assumed balanced \( (n_1 = n_2 = m) \), but irregular, and furthermore different across the two samples. Specifically, it is assumed that \( \{t_{1ij} : 1 \leq i \leq n_1, 1 \leq j \leq m_1\} \) are iid realizations from \textit{Uniform}(0,1), and \( \{t_{2ij} : 1 \leq i \leq n_2, 1 \leq j \leq m_2\} \) are iid realizations from the distribution with density \( 0.8 + 0.4t \) for \( 0 \leq t \leq 1 \). Two scenarios are considered: i) \( m = 20 \) points per curve, and ii) \( m = 100 \) points per curve. Notice this is an example where different basis function expansions are used for \( X_{1i}(\cdot) \) and \( X_{2i}(\cdot) \) for \( \delta > 0 \).

The null hypothesis that the underlying distribution is identical in the two samples is tested using CVM (Hall and Van Keilegom, 2007) and FAD testing procedures for various values of \( \delta \). Figure 2.2 illustrates the comparison between the approaches for significance level \( \alpha = 0.05 \); the results are based on 500 Monte Carlo replications. The CVM test is conducted using the procedure described in Hall and Van Keilegom (2007), and the p-value is determined based on
250 bootstrap replicates; the results are obtained using the \texttt{R} code provided by the authors. To apply marginal FPCA, and determine the marginal eigenbasis,, we use \texttt{refund} package (Crainiceanu et al., 2012) in \texttt{R}, which requires that the data are formatted corresponding to a common grid of points, with possible missingness. Thus, a pre-processing step is necessary. For each scenario, we consider a common grid of \( m \) equally spaced points in \([0, 1]\) and bin the data of each curve according to this grid. This procedure introduces missingness for the points where no data are observed. We note that this pre-processing step is not necessary if one uses \texttt{PACE} package (Yao et al., 2005a) in \texttt{Matlab}, for example. However, our preference for using open-source software, motivates the use of \texttt{refund}. Comparison of \texttt{refund} and \texttt{PACE} revealed that the two methods lead to similar results when smoothing trajectories from noisy and sparsely observed ‘functional’ data.

As Figure 2.2 illustrates, both procedures maintain the desired level of significance and the number of observations per curve, \( m_1 = m_2 \), do not seem to strongly impact the results. However, the empirical power of the FAD test increases at a faster rate than the CVM test (Hall and Van Keilegom, 2007) under all the settings considered. This should not seem surprising, as by representing the data using the marginal eigenbasis expansion as detailed in Section 2.3 we remove extraneous components. In contrast, the CVM test attempts to estimate all basis functions by smoothing the data. This can introduce error that can ultimately lower the power of the test. Additionally, due to the usage of bootstrapping to approximate the null distribution of the test, the CVM test has a much higher computational burden than the FAD.
2.5 Diffusion Tensor Image Data Analysis

The motivating application is a DTI tractography study of MS patients. The data for our study comprises 162 subjects with MS and 42 healthy controls observed at one visit. Details of this study have been described previously by Greven et al. (2011), Goldsmith et al. (2011), and Staicu et al. (2012). Parallel diffusivity and fractional anisotropy measurements are recorded at 93 locations along the corpus callosum (CCA) tract - the largest brain tract that connects the two cerebral hemispheres. Tracts are registered between subjects using standard biological landmarks identified by an experienced neuroradiologist. For illustration, Figure 2.3 displays the parallel diffusivity and fractional anisotropy profiles along the CCA for both healthy controls and MS patients. Part of this data set is available in the R-package refund (Crainiceanu et al., 2012). Our objective is to study if parallel diffusivity or fractional anisotropy along the CCA tract have the same distribution for subjects affected by MS and for controls. Such assessment would provide researchers with valuable information about whether either of these modalities, along this tract, are useful in determining axonal disruption in MS. Visual inspection of the data (see Figure 2.3) reveals that the average of fractional anisotropy seems to be different in cases than controls. It appears that for fixed tract location, parallel diffusivity exhibits a distribution with shape characteristics that depend on the particular tract location. Furthermore, the location-varying shape characteristics seem to be different in the MS versus control groups. Staicu et al. (2012) proposed a modeling approach that accounts for the features of the pointwise distribution of the parallel diffusivity in the two groups. However, they did not formally investigate whether the distribution of this DTI modality is different for the two groups. Here

![Figure 2.3: Top: Fractional anisotropy for cases and controls with the pointwise mean in red. Bottom: Parallel diffusivity for cases and controls with the pointwise mean in red.](image-url)
we apply the proposed marginal FPCA and two-sample hypothesis testing to assess whether the true distribution of parallel diffusivity and fractional anisotropy respectively, along the CCA tract is the same for MS and controls.

The parallel diffusivity and fractional anisotropy profiles are typically sampled on a regular grid (93 equal spaced locations); however, some MS patients have missing data. Additionally, the observations are assumed to be contaminated by noise. We use the marginal FPCA method discussed in Section 2.3, corresponding to sparse sampling design. The marginal mean function is estimated using penalized splines with 10 basis functions. The functional principal component decomposition was performed using the \texttt{fpea.sc} function in the \texttt{R} package \texttt{refund} and by setting the percentage of explained variance parameter to $\tau = 95\%$ (Crainiceanu et al. (2012)).

![Eigenfunction 1](image1)

![Eigenfunction 2](image2)

![Eigenfunction 3](image3)

Figure 2.4: Parallel Diffusivity (LO). Top: First three estimated eigenfunctions that explain 91% of the variation in the combined data set. Five eigenfunctions explain over 95% of the marginal variation in the data. Bottom: Box plots of the first three estimated basis coefficients.

For the parallel diffusivity data set, Figure 2.4 displays the three (of five) leading marginal eigenfunctions, along with the box plots of the corresponding coefficients presented separately for the MS and control groups. The first, second, and third functional principal component functions explain 76%, 8%, and 7% of the total variability of the combined data, respectively. The first marginal functional principal component is negative and has a concave shape with a dip around location 60 of the CCA tract. This component gives the direction along which the two curves differ the most. Near location 60 the distribution of the parallel diffusivity is highly skewed for the MS group, but not as skewed in the control group. Examination of the boxplot of the coefficients corresponding to the first eigenfunction (left, bottom panel of Figure 2.4) shows that most healthy individuals (controls) are loaded positively on this component,
yielding parallel diffusivity profiles that are lower than the overall average profile. Half of the MS subjects are loaded negatively on this component resulting in increased parallel diffusivity.

The marginal FPCA procedure estimates that five eigenfunctions account for 95% of the total variation in the parallel diffusivity data. We apply the FAD testing procedure to study whether the distributions of the five coefficients is the same for MS and controls. The p-values of the univariate tests are $p_1 = 0.00001$, $p_2 = 0.07165$, $p_3 = 0.02479$, $p_4 = 0.19887$, and $p_5 = 0.29161$; the p-value of the FAD test is thus $p = 5 \times \min_{1 \leq k \leq 5} p_k = 0.00005$. This shows significant evidence that the parallel diffusivity has different distribution in MS subjects than controls.

Figure 2.5: Fractional Anisotropy (FA). Top: First four estimated eigenfunctions that explain over 90% of the total variation in the combined FA data set. Six eigenfunctions explain over 95% of the marginal variation in the data. Bottom: Box plots of the first four estimated marginal basis coefficients.

We turn next to the analysis of the fractional anisotropy in controls and MS cases. Six functional principal components are selected to explain 95% of the total variation of the combined data. Figure 2.5 illustrates the leading four marginal eigenfunctions of the combined sample (which explain about 90% of the entire variability), along with the boxplots of the distributions of the corresponding controls/cases coefficients. The estimated first eigenfunction implies that the two samples differ in the mean function. Using the p-values of the six univariate tests are $p_1 \approx 0$, $p_2 = 0.36357$, $p_3 = 0.00833$, $p_4 = 0.58662$, $p_5 = 0.49269$, and $p_6 = 0.21427$; the p-value
of the FAD test is thus $p = 6 \times \min_{1 \leq k \leq 6} p_k \approx 0$. This shows significant evidence that the FA has different distribution in MS subjects than controls.
Chapter 3

Scan-Stratified Case Control Sampling for Modeling Blood-Brain Barrier Integrity in Multiple Sclerosis

3.1 Background and Motivation

This chapter proposes two novel methods for analysis of this data. First, using only pre-contrast images and historical information, we model probability that a given voxel in an MRI would enhance if the patient had been given a contrast agent. The model is used to assess disease activity through BBB integrity using MRI without contrast. The methodological developments are motivated by longitudinal data from an ongoing observational study at the National Institute of Neurological Disorders and Stroke (NINDS). In this work, we study a subset of high-resolution structural MRIs acquired in hundreds of patients with MS on a monthly basis with no specified end date. Each image has over one million measurements within the brain matter; this data set is growing at a rapid pace and computationally scalable methods for fitting statistical models are crucial. Our second contribution is the proposal of a novel sub-sampling technique that greatly reduces the computational burden of the estimation procedure. The proposed model, which utilizes historical information, shows superior prediction performance in terms of the receiver operating characteristic (ROC) curve when compared to the model that does not use this information. Furthermore, using a sample of approximately 750 voxels from each image yields a comparable ROC curve to the case when the proposed model is fit on the full data set that consists of one million voxels per image.
The ground truth or ‘gold standard’ for identifying enhancing white matter lesions involves the manual comparison of pre- and post-contrast T1-weighted imaging conducted by an expert neuroradiologist (Simon et al., 2006; Garcia-Lorenzo et al., 2008; Llado et al., 2012a,b). However, there are image features visible on high-resolution scans acquired on 7 tesla scanners that correlate with BBB disruption (Gaitan et al., 2012; Absinta et al., 2013). Furthermore, Shinohara et al. (2012) showed that in some cases the adjudication of BBB integrity does not require post-contrast imaging. They used a voxel-level logistic regression model for predicting enhancement using pre-contrast voxel intensities in T1w and T2-weighted (T2w) images along with the interaction between the intensities. To reduce the number of false positives, the model was fit on a subset of the voxels, defined by thresholding the T2w fluid attenuated inversion recovery images (FLAIR) images within each scan, as FLAIR images are known to have high values for lesion voxels. Taking only regions with FLAIR intensity in the top 1% provides a good set of voxels that is likely to include most enhancing lesions. Sensitivity analysis showed that the method was robust to changes in this threshold value Shinohara et al. (2012).

While thresholding on the FLAIR images does significantly reduce the sample size of the data, one criticism of this method is that all of the enhancing voxels may not be included in this sub-sample. However, in the case of MRI data of patients with MS, enhancement is rare; there are many fewer voxels that enhance compared to those that do not. It is important for our model that these rare events are all contained in the sub-sample. One can think of the data set as case-control data such that enhancing voxels are labeled as cases and non-enhancing voxels are labeled as controls. Using this setting we propose a novel sub-sampling method that takes advantage of well established theory from the case-control study literature (Prentice and Pyke, 1979; King and Zeng, 2001).

It is known that voxels belonging to new MS lesions are more likely to enhance than voxels in older lesions (Gaitan et al., 2011). Sweeney et al. (2013c) recently proposed Subtraction Based Logistic Inference for Modeling and Estimation (SuBLIME) to estimate the probability that a given voxel is part of a new or enlarging lesion between two study visits. The model relies on the historical study of patients, as the differences between images taken from the same patient at different visits are used as covariates. Then, for any given voxel in an image, one can obtain a corresponding probability that it belongs to an incident or enlarging lesion. Therefore, to predict enhancement in a contrast-free MRI study, we propose using SuBLIME voxel probabilities along with the T1w and T2w voxel intensities in a logistic regression model. This allows us to model the probability that voxels will enhance using only pre-contrast images and temporal information about lesion incidence and enlargement.

Section 3.3 presents the proposed model and compares the performance to that of Shinohara et al. (2012). Section 3.4 then proposes how the MR images acquired as part of the NINDS
study can be analyzed in a computationally scalable fashion as case-control data where the cases (enhancing voxels) are rare events. Results from the proposed methods are evaluated using cross-validation on 15 patients who received structural MR scans at 2-8 visits. In Section 3.5, the results of the new model and sampling scheme are compared to those of Shinohara et al. (2012), and we find that the proposed methodology improves prediction and reduces computational burden by reducing the number of voxels the model fits on.

3.2 NINDS Data

The MRI modalities of interest in this paper consist of FLAIR, pre-contrast T1w, and pre-contrast T2w images. For all voxels in the brain, the imaging modalities are normalized by subtracting their mean and dividing by their standard deviation (Shinohara et al., 2011a; Sweeney et al., 2013b; Shinohara et al., 2014). All images were preprocessed using the procedure outlined in the Experimental Methods section of Shinohara et al. (2012) which is outlined in the supplementary material of this paper.

For the purpose of modeling, we also obtain SuBLIME maps that provide the probability that each voxel in the brain of a patient is part of a new or enlarging lesion since the patient’s last study visit (Sweeney et al., 2013c). The probabilities are estimated by assuming a logistic model in which the covariates are T1w, T2w, proton density-weighted (PDw), and FLAIR normalized intensities along with the lag time between consecutive visits, the relative change in these imaging modalities, and their interactions with time. SuBLIME maps have not previously been used as historical covariates in the context of image analysis. More details about estimation of these maps are provided in the supplementary material of this paper.

Additionally, we obtain binary radiologist manual segmentation masks (RGS) which we consider to be the gold standard of indicating where enhancing lesions exist, and brain masks indicating where cerebral tissue lies in each image. In order to build the proposed predictive model, we use these RGS masks as the response variable. The masks are created by a neuroradiologist (D.S.R.) with over 10 years of clinical experience with MR imaging in MS. To define these masks, the neuroradiologist uses all available post-contrast imaging modalities including, but not limited to, the post-contrast T1w images. The details of the pre-processing of the images in this study are described in appendix A.2. For the analysis presented in this paper we use 77 brain MRIs from 15 individuals with MS, obtained under an institutional review board-approved protocol. Each of these patients was scanned between 2 and 9 times. Informed consent was obtained from all participants.
3.3 Logistic Enhancement Models

We present a model for predicting voxel enhancement using pre-contrast MRIs by advancing the work of Shinohara et al. (2012) and Sweeney et al. (2013c). Shinohara et al. (2012) showed that voxel enhancement can be predicted using a cross-sectional logistic model that is fit using the pre-contrast T1w and T2w voxel intensities along with the interaction under the assumptions of independence between scans and voxels. We define the incident of enhancement using a binary variable, $E_i(v,t_{ij})$ which equals 1 if voxel $v$ in the brain of patient $i$ enhances at time $t_{ij}$ and 0 otherwise. The model presented in Shinohara et al. (2012) assumes that enhancement across images for patient $i$ at time $t_{ij}$ are independent so that,

$$\logit\left\{PrE_i(v,t_{ij})\right\} = \beta_0 + \beta_1 M_{i,1}(v,t_{ij}) + \beta_2 M_{i,2}(v,t_{ij}) + \beta_{12} M_{i,1}(v,t_{ij}) M_{i,2}(v,t_{ij})$$

(3.1)

where intensity on imaging modality $k$ in voxel $v$ for subject $i$ at time $t_{ij}$ is denoted as $M_{i,k}(v,t_{ij})$ for $i = 1 : n$, $j = 1 : m_{i,k}$, and $m_{i,k}$ is the number of scans using modality $k$ for subject $i$. Here, $M_{i,1}$ and $M_{i,2}$ refer to the T1w and T2w images for patient $i$ respectively.

In Model 3.1, the fact that new or enlarging MS lesions are more likely to enhance (Gaitan et al., 2011) is not taken into account. It is our assertion that one may include a covariate that describes lesion age to improve predictive ability. This information can be obtained by having a radiologist manually indicate where new or enlarging lesions exist in an image. However, manual identification is extremely time consuming and prone to errors even for the most expert neuroradiologists in this field. Additionally, the process is subject to both intra-observer and inter-observer variability (Llado et al., 2012a). Therefore, we propose the use of the SuBLIME method to estimate the probability that a voxel belongs to a new or enlarging lesion (Sweeney et al., 2013c).

One of the main contributions of this paper is the development of a predictive model that uses information about temporal lesion behavior to predict the probability of a voxel enhancing. The proposed modeling procedure is a two-step process. First, for each voxel used for fitting the model, we acquire historical information about whether the voxel belongs to a new or enlarging lesion. This can either be obtained from a mask developed by a radiologist or from the SuBLIME method (Sweeney et al., 2013c) which estimates the corresponding probabilities. In this paper, we employ the latter technique. For this, we define the indicator of incidence as a binary variable, $W_i(v,t_{ij})$, which is equal to one if subject $i$ has new lesion incidence in voxel $v$ at time $t_{ij}$ and zero otherwise. The SuBLIME probability maps, $SP_i(v,t_{ij}) := Pr[W_i(v,t_{ij}) = 1]$, are the probabilities that voxel $v$ is with a new or enlarging lesion at time $t_{ij}$. As described in the previous section, these maps are obtained using longitudinal study information. Details about the estimation of this covariate using SuBLIME Sweeney et al. (2013c) is presented in appendix A.1.

The second step of our fitting procedure consists of modeling the probability that a voxel
enhances by accounting for T1w, T2w pre-contrast images, their interaction, and the covariate describing the new or enlarging lesion incidence. Specifically we assume the following logistic model,

$$\text{logit}\{\Pr(E_i(v, t_{ij}))\} = \beta_0 + \beta_1 M_{i,1}(v, t_{ij}) + \beta_2 M_{i,2}(v, t_{ij}) + \beta_3 S_{Pi}(v, t_{ij}) + \beta_{12} M_{i,1}(v, t_{ij}) M_{i,2}(v, t_{ij}).$$ (3.2)

We compare the predictions of Model (3.1) and Model (3.2) by analyzing the classification performance using the empirical receiver operating characteristic (ROC) curve. We consider both the full and partial AUC of the ROC curves, obtained through cross-validation. We also investigate the impact of using different sub-sampling techniques that aim to reduce the computational expense and improve prediction. In the next section, we present a modified case-control sampling technique, which we show to further improve the predictions.

### 3.4 Sampling on 3D Magnetic Resonance Images

Our data consist of images that each contain over one million voxels, so the data set analyzed in this paper contains over 77 million voxels. Furthermore, in this data set, the average proportion of voxels that are known to enhance over all RGS masks is 0.004, indicating that enhancement is an extremely rare event in these data. Developing a computationally inexpensive fitting procedure would thus be very beneficial for researchers. However, it is also important that this procedure takes into account the rarity of enhancement. In this section we propose a procedure that not only accounts for this, but reduces the computational time by two orders of magnitude and maintains high predictive power. The sub-sampling method will facilitate large scale analysis of the proposed predictive model in future studies.

We begin by considering the population of interest to be all voxels in all brains of people with MS. The full data set under study consists of the voxels in the brain scans of the subjects in the observational study. Using case-control terminology, we refer to the voxels that are identified (by the RGS) as enhancing and non-enhancing by cases and controls respectively.

Sub-sampling without replacement from the total number of voxels would often result in a sample that only contains controls. Therefore, in order to ensure full representation of the cases and maintain some of the spatial dependence in the sampling scheme, we propose a scan-stratified case-control (SSCC) sampling approach. Let $RGS_i(v, t_{ij})$ be the binary RGS mask, which equals 1 if voxel $v$ in the brain image of patient $i$ taken at time $t_{ij}$ is defined as enhancing and 0 otherwise. Denote the mask over all voxels $v$ in the brain image of patient $i$ taken at time $t_{ij}$ as $RGS_{ij}$. Furthermore, let $l_{ij}$ be the number enhancing voxels in $RGS_{ij}$ and $\bar{n}_{e}$ be the average number of enhancing voxels over all masks. Then for each image taken of patient $i$ at time $t_{ij}$, the sub-sample of cases consists of all enhancing voxels as defined by the $RGS_{ij}$. The
control sample is constructed by taking a simple random sample from all non-enhancing voxels in each scan with the following rules:

- If \( l_{ij} > 0 \) (the RGS contains enhancing voxels): take a simple random sample of size five times \( l_{ij} \) from the non-enhancing voxels identified by \( RGS_{ij} \).

- If \( l_{ij} = 0 \) (the RGS does not contain any enhancing voxels): take a simple random sample of size five times \( \bar{n}_e \) from the non-enhancing voxels identified by \( RGS_{ij} \).

The constant five is recommended in King and Zeng (2001) for the case of fitting a logistic model to a case control sample for rare event data. Let \( SSCC_{ij} \) be the set of all voxels in the case-control sub-sample for patient \( i \) at time \( t_{ij} \) (determined using \( RGS_{ij} \)). Then the final case-control sub-sample of voxels used to fit the model is \( SSCC = \bigcup_{ij} SSCC_{ij} \).

The SSCC sampling method ensures that all of the cases are included in the sub-sample used to fit logistic models (1) and (2). This is desirable because including enough cases in the sample can reduce the variation of logistic regression coefficient estimates (Imbens, 1992; Cosslett, 1981; Imbens and Lancaster, 1996; King and Zeng, 2001). However, if only voxels \( v \in SSCC \) are used to fit the proposed models, \( \hat{\beta}_1, \hat{\beta}_2, \) and \( \hat{\beta}_{12} \) will be consistent, but \( \hat{\beta}_0 \) may be biased (Prentice and Pyke, 1979; King and Zeng, 2001). Since our goal is to maintain high predictive power when only using \( \{v : v \in SSCC\} \) to fit models (1) and (2) it is important to address this problem.

The bias in this estimate can be easily removed by using the correction presented in King and Zeng (2001). In their paper, they focus on model estimation when fitting logistic models with a case-control sample for rare event data. However, they are motivated by the difficulty of obtaining large amounts of data. While it is not difficult for us to obtain a large amount of control data, it does greatly increase the computational expense of running the model with a large number of scans. Therefore, the goal is the same in both scenarios; to use a case-control sample to fit a logistic model without loss of prediction power.

Denote the intercept estimate from fitting a logistic model using only \( \{v : v \in SSCC\} \) as \( \hat{\beta}_0^{SSCC} \), where the super-script specifies the SSCC sub-sample used for fitting the model. Let \( \tau \) be the population proportion of cases and \( \bar{y} \) be the proportion of cases in the sample. Then the bias-corrected coefficient estimate derived in King and Zeng (2001) is, \( \hat{\beta}_0 = \hat{\beta}_0^{SSCC} - \ln [(1 - \hat{\tau}/\hat{\tau}) (\bar{y}/1 - \bar{y})] \). We obtain \( \hat{\tau} \) by computing the proportion of enhancing voxels in \( \bigcup_{ij} RGS_{ij} \). The number of enhancing voxels in \( SSCC \) is computed to obtain \( \bar{y} \). In the next section, we show that using the proposed SSCC sampling method along with this bias correction allows us to maintain powerful predictive power while drastically decreasing the computational time for fitting the models.
3.5 Enhancement Prediction Results

The full data set analyzed in this paper consists of 77 million voxels. Using the FLAIR sub-sampling method presented by Shinohara et al. (2012) reduces the sample size to under 5.6 million voxels. The proposed SSCC sub-sampling method further reduces the sample size to 58 thousand voxels. In this section we compare the predictive performance of both models (1) and (2) when fitting on the full data set, the FLAIR sample, and the SSCC sample. The estimated ROC curves are used to determine which of the six ($2 \times 3$) possible model-sampling technique combinations achieves the best classification in terms of the false positive rate (FPR) and true positive rate (TPR) trade-off.

The ROC curves are estimated using cross-validation by resampling subjects. To accomplish this, we split the 15 subjects into two sets. All images from eight randomly selected subjects are placed in a fitting set and all images from the remaining seven subjects are placed in the validation set. The estimated ROC curve for each model-sample combination is taken to be the average ROC calculated on the validation set across 100 bootstrap replications. We conducted a sensitivity analysis (omitted) to investigate the stability of the average ROC curves, and we found 100 replications sufficient. To ensure proper comparison of the prediction performance, we fit the models using each of the three sets of data (all voxels, 1% FLAIR, SSCC), but validate using all brain voxels.

Due to the rarity of enhancement, analyzing FPRs larger than the actual event rate is not of interest. If the event rate is low, as in this data set, then investigating the trade-off of FPR and TPR, for FPRs much greater than the event rate, does not give us information about the prediction on this or similar data sets. To address this, we truncate the ROC curve at a fixed FPR, a common method of evaluating classification methods for rare event data (Pepe et al., 2013). To find a reasonable cutoff for such a partial ROC analysis, we first note that in our data set the average rate of enhancing voxels in the entire brain over all scans is 4 per 1000. Since the data consist of 3D images, we consider each voxel in the context of a $3 \times 3 \times 3$ cube surrounding it; by multiplying .004 with $3^3 = 27$, we thus deem FPRs above 0.108 irrelevant.

Figure 3.1 provides plots of the partial ROC (pROC) curves comparing models (1) and (2) when fit using each of the three data sub-sets (full data, 1% FLAIR, and SSCC). The partial area under the curve (pAUC) is a useful summary of the pROC reported in Table 3.1. This value is computed by taking the area under the partial curve and scaling it to lie between 0 and 1. Note that model (2) outperforms model (1) when each is fit on all three data sets.

We observe that SSCC and full-brain sampling perform comparably for both methods, and both outperform FLAIR sampling. This can also be observed in Figure 3.2 which overlays the pROC curves corresponding to the three data sets for each model separately. From this pROC analysis, we conclude that i) adding the SuBLIME covariate does seem to improve model (1);
Figure 3.1: Partial ROC curves comparing Models (1) and (2) (blue line (circles) and solid red line (squares) respectively) for: A-All voxels; B-Top 1% FLAIR sub-sampling; and C-SSCC sub-sampling. The dashed line in all curves corresponds to having TPR=FPR.

and ii) fitting the models using the SSCC sampling method maintains predictive power while vastly improving computational efficiency. Using a case-control sample we can find estimates for model (2) in 1.6 seconds using SSCC compared to 286.8 seconds using all voxels in the full data set.

An example of the improvement in classification of model (2) is presented in Figure 3.3 which displays an axial slice from one of patient’s images used in this analysis. Figures 3.3A

Table 3.1: Partial AUC estimates and corresponding 95% confidence intervals (CI) for curves in figures 3.1 and 3.2.

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th></th>
<th>Model 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>pAUC</td>
<td>95% CI</td>
<td>pAUC</td>
<td>95% CI</td>
</tr>
<tr>
<td>All Voxels</td>
<td>0.52</td>
<td>(0.49,0.55)</td>
<td>0.74</td>
<td>(0.72,0.76)</td>
</tr>
<tr>
<td>1% FLAIR</td>
<td>0.29</td>
<td>(0.27,0.32)</td>
<td>0.64</td>
<td>(0.61,0.67)</td>
</tr>
<tr>
<td>SSCC</td>
<td>0.53</td>
<td>(0.50,0.56)</td>
<td>0.78</td>
<td>(0.76,0.80)</td>
</tr>
</tbody>
</table>

Figure 3.2: Partial ROCs curves from using the three data sets under study to fit model (1) (A) and model (2) (B).
and 3.3B display the pre-contrast T1w and T2w images respectively. Figures 3.3C and 3.3D display the post-contrast FLAIR and T1w images respectively. The RGS which displays the location of the lesion defined by the neuroradiologist is shown in Figure 3.3H. Additionally, we present the resulting probability maps from models (1) and (2). The lesion (red) in the upper right portion of the images is detected in Figure 3F but not in Figure 3E, which display the predicted probabilities from model (2) and (1) respectively. However, it is identified as a newly enhancing lesion by the SuBLIME model which is displayed in Figure 3D. Note that the model (1) probability map (Figure 3E) also shows the ventricular CSF as enhancing lesion, indicating extrapolation errors outside of the FLAIR-hyperintense voxels, whereas model (2) probability map (Figure 3F) does not exhibit these distracting artifacts. Shinohara et al. (2012) did not observe this artifact as they only predicted on the set of voxels corresponding to the top 1% of FLAIR histogram. We see this in our analysis because we predict on all voxels containing brain matter. Reducing this type of artifact in the predictions further motivates the use of the SSCC sampling method along with model (2). The coefficient estimates for the model-sample combinations are provided in Table 3.2.
Table 3.2: Coefficient estimates when model 1 (left) and model 2 (right) are fit using all voxels in the corresponding sets.

<table>
<thead>
<tr>
<th>Model 1: All Voxels</th>
<th>Estimate</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>-10.18</td>
<td>(-10.21, -10.14)</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>1.08</td>
<td>(1.03, 1.13)</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>1.98</td>
<td>(1.96, 2.00)</td>
</tr>
<tr>
<td>( \beta_{12} )</td>
<td>0.88</td>
<td>(0.85, 0.90)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FLAIR 1%</th>
<th>Estimate</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>-7.81</td>
<td>(-7.85, -7.77)</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>0.01</td>
<td>(-0.06, 0.09)</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>2.00</td>
<td>(1.98, 2.04)</td>
</tr>
<tr>
<td>( \beta_{12} )</td>
<td>0.52</td>
<td>(0.48, 0.56)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SSCC</th>
<th>Estimate</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>-10.79</td>
<td>(-10.85, -10.74)</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>2.02</td>
<td>(1.93, 2.11)</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>2.88</td>
<td>(2.82, 2.94)</td>
</tr>
<tr>
<td>( \beta_{12} )</td>
<td>1.32</td>
<td>(1.27, 1.37)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model 2: All Voxels</th>
<th>Estimate</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>-10.23</td>
<td>(-10.27, -10.20)</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>0.814</td>
<td>(0.76, 0.87)</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>1.53</td>
<td>(1.50, 1.56)</td>
</tr>
<tr>
<td>( \beta_3 )</td>
<td>10.03</td>
<td>(9.94, 10.12)</td>
</tr>
<tr>
<td>( \beta_{12} )</td>
<td>0.95</td>
<td>(0.92, 0.98)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FLAIR 1%</th>
<th>Estimate</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>-7.65</td>
<td>(-7.69, -7.60)</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>-0.51</td>
<td>(-0.59, -0.43)</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>1.40</td>
<td>(1.37, 1.44)</td>
</tr>
<tr>
<td>( \beta_3 )</td>
<td>6.74</td>
<td>(6.64, 6.83)</td>
</tr>
<tr>
<td>( \beta_{12} )</td>
<td>0.97</td>
<td>(0.92, 1.02)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SSCC</th>
<th>Estimate</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>-11.42</td>
<td>(-11.49, -11.35)</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>1.94</td>
<td>(1.83, 2.06)</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>2.53</td>
<td>(2.46, 2.60)</td>
</tr>
<tr>
<td>( \beta_3 )</td>
<td>37.09</td>
<td>(34.93, 39.27)</td>
</tr>
<tr>
<td>( \beta_{12} )</td>
<td>1.09</td>
<td>(1.03, 1.15)</td>
</tr>
</tbody>
</table>

The gain in predictive performance that we observed is unprecedented, despite the simplicity and strong parametric assumptions of the model. Since using the historical covariate yields such an improvement, future work will focus on longitudinally modeling enhancement probabilities.
Chapter 4

A Lag Functional Linear Model for Prediction of Magnetization Transfer Ratio in Multiple Sclerosis Lesions

4.1 Introduction

Multiple sclerosis (MS) is an inflammatory disease of the central nervous system in which the myelin sheaths around the axons of the neurons in the brain and spinal cord are damaged. Lesions in the brain and spinal cord are known to characterize this disease. Magnetic resonance imaging (MRI) is sensitive to these lesions and is often used by clinicians to evaluate these abnormalities. One benchmark diagnostic criterion that is observed via MRI is the number of white matter lesions (Polman et al., 2005; McDonald et al., 2001). Conventional MRI modalities that are used to observe lesions are T1-weighted (T1w), T2-weighted, and T2-weighted fluid attenuated inversion recovery (FLAIR) imaging. However, it has been argued that these rudimentary measures of lesion burden, including the number and volume of lesions, only moderately correlate with disability (Hawkins et al., 1990; Brex et al., 2002; Schmierer et al., 2004). Therefore, much research in the field has centered on investigating the use of images that directly quantify tissue damage and are believed to correlate with disability such as magnetic transfer ratio (MTR) maps. Lesions can result from various etiologies including stroke, diabetes, HIV, and aging. In all of these areas, the investigation of these lesions in terms of tissue damage has been a primary objective of many neuroimaging studies.

Our goal is to estimate these quantitative images within lesions using only commonly acquired T1w and FLAIR images. This will give us insight into how these images enrich our understanding of the processes observed in lesions and allow clinicians to estimate MTR at
past or future study visits when these modalities are not measured. For this purpose, we propose a voxel-level model that uses longitudinal measurements of the T1w and FLAIR images over a fixed window of time to predict MTR within lesions. The model yields improvement over cross-sectional models which do not account for temporal patterns of lesional development and repair. This will ultimately allow researchers to study the effect of treatments on tissue damage in patients with MS.

We analyze data that were obtained as part of a natural history study of MS patients at the National Institute of Neurological Disorders and Stroke (NINDS). Various MRI modalities were acquired for a group of 53 MS patients. Each patient was scanned between 9 and 38 times over a period of 5.5 years. In this longitudinal study, the number of visits $m_i$ for $i \in \{1, \ldots, 53\}$ varies for each patient, and MTR maps, T1w, and FLAIR images were acquired at each of the visits. For our analysis, the voxels in all of the images are temporally registered to obtain voxel-level intensity trajectories over time. The images are also segmented to obtain the set of voxel trajectories that correspond to areas of white matter containing MS lesions. Our objective is to relate current MTR intensity in a given voxel (within a lesion) to the T1w and FLAIR intensities in the same voxel observed over time. The observed voxel-level trajectories are discrete observations arising from a continuous process over time. Therefore, to predict MTR we employ a function-on-function regression framework.

We propose a model that facilitates prediction of the functional response (MTR), using only a specified fixed window (six months) of history of the functional predictors (T1w and FLAIR) at the voxel-level. As documented by Van Den Elskamp et al. (2008), progression of tissue damage within a lesion is primarily observed, using MRI, for six months. We will develop a model that will use only the intensity of the T1w and FLAIR images observed in the previous six months of historical data to predict MTR at the current time point. Current methods in the functional data analysis literature are not able to model the relationship between a functional response and functional predictors when it is assumed that only the recent past of the predictors affect the current value of the response.

Due to a surge of applications for functional data, there have been several methodological and theoretical advancements in the area; see Besse and Ramsay (1986); Rice and Silverman (1991); Ramsay and Silverman (2005); Ferraty et al. (2007); Horváth and Kokoszka (2012) to name a few. To estimate a functional response using the value of the functional predictor at the current point, the functional concurrent model was proposed by Hastie and Tibshirani (1993). Then, to use the entire range of one functional predictor to predict a functional response Ramsay and Silverman (1997) developed the functional linear model.

For the scientific problem of interest in this chapter, it is not appropriate to use the future of the functional covariate to estimate the functional response at a specific time point. To account
for this, Malfait and Ramsay (2003) propose a model for a functional response that uses only the history of one functional predictor. In their model, they represent the coefficient function using a triangular basis function that is estimated using information from the predictor and estimate a coefficient function at each observation point. Kim et al. (2011) propose a least-squares approach to fitting this model by estimating the coefficient function at each timepoint.

Until recently, the work related to function-on-function regression has only been able to account for one functional predictor; see Ramsay and Dalzell (1991); Ramsay and Silverman (1997); Yao et al. (2005b); He et al. (2010); Wu et al. (2010). For a complete review of the current literature in this area, see Morris et al. (2003). Recently, Ivanescu et al. (2014) developed a penalized regression method for association models between functional responses and multiple functional predictors, where the current response is allowed to depend on the entire trajectory of the predictor. In this chapter we continue the study of regression models for a functional response and multiple functional predictors when response depends solely on a fixed window of the functional predictors. The main innovation of our work is to propose a functional linear model to estimate the current value of MTR using only the last six months of both the T1w and FLAIR voxel intensities. To estimate the regression parameter functions, we propose two methods: 1) a semi-local smoothing approach; and 2) a global smoothing approach.

The rest of the chapter is organized as follows: we introduce the modeling framework in Section 4.2. Two estimation procedures are presented in Sections 4.2.1 and 4.2.2. We evaluate the proposed methods in a simulation experiment in Section 4.3. In Section 4.4 the proposed estimation approaches are then applied to the motivating MTR data application, for predicting MTR trajectories using the recent past of T1w and FLAIR voxel-level intensities.

### 4.2 Methodology

We begin with introducing the historical functional linear model for multiple functional covariates. Suppose we observe data, \([Y_{ij}, t_{ij} : i \in \{1,...,n\} \text{ and } j \in \{1,...,m_{0i}\}], [X_{1ij}, r_{1ij} : i \in \{1,...,n\} \text{ and } j \in \{1,...,m_{0i}\}]\) and \([X_{2ij}, r_{2ij} : i \in \{1,...,n\} \text{ and } j \in \{1,...,m_{0i}\}]\), where \(t_{ij}, r_{1ij}, r_{2ij} \in \mathcal{T}\), a bounded and closed interval. Let \(Y_{ij} = Y_i(t_{ij})\), \(X_{1ij} = X_1(t_{ij})\), and \(X_{2ij} = X_2(t_{ij})\) and assume that \(Y(\cdot), X_1(\cdot)\) and \(X_2(\cdot)\) are independent and square-integrable random functions over \(\mathcal{T}\); without loss of generality assume \(E[X_{1i}(t)] = E[X_{2i}(t)] = 0\). For simplicity take \(\mathcal{T} = [0,1]\), and assume that the predictor functions are observed on a dense grid of points and without noise, and the response is observed on a dense and equidistant grid of points \(t_{ij} = t_j\). Furthermore, assume that the response \(Y_{ij}\) are observed with noise and arise from the following model,

\[
\begin{align*}
Y_{ij} &= f_i(t_j) + \epsilon_i(t_j) \\
f_i(t_j) &= \beta_0(t_j) + \int_0^{\Delta_1} \beta_1(s,t_j) X_{1i}(t_j - s)ds + \int_0^{\Delta_2} \beta_2(s,t_j) X_{2i}(t_j - s)ds
\end{align*}
\]

(4.1)
where \( i = \{1, ..., n\}, j = \{1, ..., m\} \). Let \( E[Y_i|X_{1i}(\cdot), X_{2i}(\cdot)] = f_i(t_j), \beta_0 : \mathcal{T} \to \mathbb{R} \), and \( \beta_1, \beta_2 : [0, \Delta] \times \mathcal{T} \to \mathbb{R} \). For intuition, notice that this model is equivalent to writing, \( Y_{ij} = \beta_0(t_j) + \int_{t_j-\Delta_1}^{t_j} \beta_1(t_j - s, t_j)X_{1i}(s)ds + \int_{t_j-\Delta_2}^{t_j} \beta_2(t_j - s, t_j)X_{2i}(s)ds + \epsilon_i(t_j) \). The model assumes that given the entire predictor trajectories, the response is only affected by the predictor values over recent time-windows of length \( \Delta_1 \) and \( \Delta_2 \) respectively. These lag parameters are chosen based on scientific knowledge. For simplicity of illustration we let \( \Delta = \Delta_1 = \Delta_2 \). The time-varying coefficient functions, \( \beta_1(\cdot, \cdot) \) and \( \beta_2(\cdot, \cdot) \), weigh the predictor values over the \( \Delta \) lag-window, and quantify the effect of the predictors on \( Y_i(t) \) for \( s \in [t - \Delta, t] \). The measurement error \( \epsilon_i(\cdot) \) is a zero-mean random process that is assumed independent of the predictors. We refer to this model as the multiple covariate lag Functional Linear Model (lagFLM). Model (4.1) extends Malfait and Ramsay (2003) and Kim et al. (2011) to account for multiple functional covariates.

As is common in the functional data analysis literature, we model the smooth effects using basis function expansions. Specifically, let \( \{B_{s1,k}(s)\}_k \) and \( \{B_{s2,k}(s)\}_k \) be two pre-specified bases functions on \( [0, \Delta] \). Then \( \beta_1(s,t) \) can be represented as \( \beta_1(s,t) = \sum_{k=1}^{K_1} B_{s1,k}(s)a_{1k}(t) \), where \( K_1 \) is a large enough to capture the flexibility of the model and the \( a_{1k}(t)'s \) are unknown time-varying coefficients functions defined on \( \mathcal{T} \). We use a similar basis expansion for \( \beta_2(s,t) = \sum_{k=1}^{K_2} B_{s2,k}(s)a_{2k}(t) \) where \( a_{2k}(t)'s \) are unknown time-varying coefficients functions defined on \( \mathcal{T} \). Without loss of generality we assume \( K_1 = K_2 = K \), so that substituting these sums into equation (4.1) yields,

\[
f_i(t_j) = \beta_0(t_j) + \sum_{k=1}^{K} a_{1k}(t_j) \int_0^{\Delta} X_{1i}(t_j - s)B_{s1,k}(s)ds + \sum_{k=1}^{K} a_{2k}(t_j) \int_0^{\Delta} X_{2i}(t_j - s)B_{s2,k}(s)ds.
\]

Furthermore, by defining \( \tilde{X}_{1k,i}(t_j) = \int_0^{\Delta} X_{1i}(t_j - s)B_{s1,k}(s)ds \) and \( \tilde{X}_{2k,i} \) similarly, equation (4.2) can be reduced to

\[
f_i(t_j) = \beta_0(t_j) + \sum_{k=1}^{K} \tilde{X}_{1k,i}(t_j)a_{1k}(t_j) + \sum_{k=1}^{K} \tilde{X}_{2k,i}(t_j)a_{2k}(t_j).
\]

We discuss two approaches for estimating these model parameters. The first approach, abbreviated as PW considers, a fitting of (4.3) pointwise (hence PW) for each observed time point \( t_j \); the smoothness of the model parameters over \( \mathcal{T} \) is controlled by using estimators which are smooth over \( \mathcal{T} \). The second approach, abbreviated GB considers basis functions expansions using pre-specified bases for each smooth model parameter and intuitively controls the smoothness over \( \mathcal{T} \) globally (hence GB). The two approaches are detailed in turn next.

### 4.2.1 PW Model Representation

In this section, we develop an estimation procedure of the lagFLM model that depends on the continuous second moments of the additive predictors. It does not require direct computation of
\( \tilde{X}_{pk,i}(t) \) and accounts for the correlation between the functional predictors. This representation of the function-on-function regression model will allow us to obtain smooth pointwise estimates for the coefficient functions. A similar approach was taken by Kim et al. (2011) for the recent history functional linear model with one functional predictor. Estimation of the coefficient functions and prediction of the response is discussed for the dense case in this section and then extended to more realistic settings.

Take a fixed value of \( t_j \), and denote by \( \alpha_{1kj} = \alpha_{1k}(t_j) \) and \( \alpha_{2kj} = \alpha_{2k}(t_j) \). The smoothness of the model parameters, \( \beta_1(s,t) \) and \( \beta_2(s,t) \), over \( s \) is controlled by the number of B-spline functions, and thus by the choice of \( K \). Our approach is to consider \( K \) to be large enough and penalize the basis coefficients. For fixed \( t \), this comes down to regularizing the coefficients \( \alpha_{1kj} \) and \( \alpha_{2kj} \) respectively. In particular, we impose a ridge constraint \( \sum_{k=1}^{K} \alpha_{1kj}^2 \leq C_1 \) and \( \sum_{k=1}^{K} \alpha_{2kj}^2 \leq C_2 \) for some constants \( C_1 \) and \( C_2 \) (Hoerl and Kennard, 1970; Hastie et al., 2009; Shao and Deng, 2012). Then for each time point \( t_j \) on the pre-specified grid at which the response is observed, we can obtain estimates of the coefficient functions by minimizing the least square error loss function, \( \sum_i (Y_{ij} - f_i(t_j))^2 \) subject to the latter constraint. This is equivalent to minimizing the following penalized criterion:

\[
\sum_i (Y_{ij} - f_i(t_j))^2 + \lambda_1 \sum_k \alpha_{1kj}^2 + \lambda_2 \sum_k \alpha_{2kj}^2,
\]

where \( \lambda_1 \) and \( \lambda_2 \) are the regularization parameters. Let \( \alpha_{pj} = [\alpha_{p1j} \cdots \alpha_{pKj}]^\top \), so the solution to the minimization of this criterion is given by

\[
\begin{bmatrix}
\hat{\alpha}_{1j} \\
\hat{\alpha}_{2j}
\end{bmatrix}
= \left( \begin{bmatrix} Z_j^\top Z_j & \lambda_1 I_K \\ \lambda_1 I_K & \lambda_2 I_K \end{bmatrix} \right)^{-1} \begin{bmatrix} Z_j^\top (Y_j - \mu_Y(t_j)) \end{bmatrix}, \quad (4.5)
\]

where \( I_K \) is the identity matrix, \( Y_j \) is the vector of all observations at time \( t_j \), and

\[
Z_j = \begin{bmatrix}
\tilde{X}_{11,1}(t_j) & \cdots & \tilde{X}_{1K,1}(t_j) & \tilde{X}_{21,1}(t_j) & \cdots & \tilde{X}_{2K,1}(t_j) \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\tilde{X}_{11,n}(t_j) & \cdots & \tilde{X}_{1K,n}(t_j) & \tilde{X}_{21,n}(t_j) & \cdots & \tilde{X}_{2K,n}(t_j)
\end{bmatrix}. \quad (4.6)
\]

For fixed \( t_j \), the analytic solution (4.5) is the solution in a ridge regression framework, for fixed values of the tuning parameters \( \lambda_1 \) and \( \lambda_2 \). In our case, \( \alpha_{1kj} \)’s and \( \alpha_{2kj} \)’s represent the evaluations of smooth functions \( \alpha_{1k}(t) \) and \( \alpha_{2k}(t) \) at times \( t = t_j \). Selecting optimal values of \( \lambda_1 \) and \( \lambda_2 \) for each \( t \) could lead to increased variability of the tuning parameters over \( t \).

One approach to accommodate this situation is to assume that the tuning parameters are constant across \( t \). The parameter functions \( \alpha_{1k}(t) \) and \( \alpha_{2k}(t) \) can be estimated using the probability limits of the standardized block submatrices, \( Z_j^\top Z_j \) and \( Z_j^\top Y_j \) that are included in expression (4.5). Specifically, for arbitrary \( t \),

\[
\begin{bmatrix}
\hat{\alpha}_1(t) \\
\hat{\alpha}_2(t)
\end{bmatrix}
= \left( \begin{bmatrix} \hat{G}_{11}(t) & \hat{G}_{12}(t) \\ \hat{G}_{21}(t) & \hat{G}_{22}(t) \end{bmatrix} + \begin{bmatrix} \frac{\lambda_1}{n} I_K & 0 \\ 0 & \frac{\lambda_2}{n} I_K \end{bmatrix} \right)^{-1} \begin{bmatrix} \hat{G}_{1Y}(t) \\ \hat{G}_{2Y}(t) \end{bmatrix}, \quad (4.7)
\]

39
where $\hat{G}_{pY}(t) = [\hat{G}_{pY_1}(t) \cdots \hat{G}_{pY_K}(t)]^\top$ is a $K \times 1$ vector, $\hat{G}_{pY_1}(t)$ is an estimator of $G_{pY_1}(t) = \text{cov} \left( \hat{X}_{pl,i}(t), Y_{ij} \right)$, $\hat{G}_{pq}(t) = [\hat{G}_{\hat{X}_{pl,i} \hat{X}_{qk}}(t)]_{kl}$ is a $K \times K$ matrix, and $\hat{G}_{\hat{X}_{pl,i} \hat{X}_{qk}}(t)$ is an estimator of $G_{\hat{X}_{pl,i} \hat{X}_{qk}}(t) = \text{cov} \left( \hat{X}_{pl,i}(t), \hat{X}_{qk}(t) \right)$ for $p, q \in \{1, 2\}$. Although the model framework and estimation is discussed for two functional predictors, one can straightforward extend it to accommodate three or more functional predictors.

To obtain the necessary quantities, observe that
cov \left( \hat{X}_{pl,i}(t_j), \hat{X}_{qk,i}(t_j) \right) = \int_0^\Delta \int_0^\Delta E \left[ X_{pi}(t_j - s_1) X_{qi}(t_j - s_2) \right] B_{sp,l}(s_1) B_{sq,k}(s_2) ds_1 ds_2
cov \left( \hat{X}_{pl,i}(t_j), Y(t_j) \right) = \int_0^\Delta G_{X_pY}(t_j - s, t_j) B_{sp,l}(s) ds \text{ where } G_{X_pY}(s_1, s_2) \text{ is the covariance between } X_{pi}(s_1) \text{ and } Y_i(s_2).

To use expression (4.7), one first needs to obtain estimates of $G_{11}(t)$, $G_{12}(t)$, $G_{21}(t)$, $G_{22}(t)$, $G_{1Y}(t)$, and $G_{2Y}(t)$. Once estimates of $\alpha_1(t)$ and $\alpha_2(t)$ are available the initial parameter functions are estimated as $\hat{\beta}_1(s, t) = \sum_{k=1}^{K_1} B_{s1,k}(s) \hat{\alpha}_{1k}(t)$ and similarly for $\hat{\beta}_2(s, t)$. The consistency of these coefficient function estimators relies on consistent estimation of the smooth covariances $G_{X_pX_q}$ and $G_{X_pY}$ as stated in proposition 1.

**Proposition 1** Assume the following:

[A1] the estimators for the functional mean and covariances are consistent, i.e.
\[
\sup_{t \in \tau} |\hat{\mu}_p(t) - \mu_p(t)| = O_p(n^{-\alpha})
\]
\[
\sup_{t \in \tau} |\hat{\mu}_Y(t) - \mu_Y(t)| = O_p(n^{-\alpha})
\]
\[
\sup_{s_1, s_2 \in \tau} |\hat{G}_{X_pX_q}(s_1, s_2) - G_{X_pX_q}(s_1, s_2)| = O_p(n^{-\alpha})
\]
\[
\sup_{s_1, s_2 \in \tau} |\hat{G}_{X_pY}(s_1, s_2) - G_{X_pY}(s_1, s_2)| = O_p(n^{-\alpha})
\]

where $0 < \alpha < 1$,

[A2] $B_{s1,k}(s)$ and $B_{s2,k}(s)$ are uniformly bounded, i.e. $\sup_{s \in [0, \Delta]} |B_{s1,k}(s)| ds \leq C < \infty$ and similarly for $B_{s2,k}(s)$,

[A3] $\lambda_1$ and $\lambda_2$ are bounded, and

[A4] the matrix
\[
G(t) = \begin{bmatrix}
G_{11}(t) & G_{12}(t) \\
G_{21}(t) & G_{22}(t)
\end{bmatrix}
\]
is invertible for all $t$.

Then the estimator $\hat{\beta}_p(s, t)$ is consistent, i.e. $\sup_{s \in [0, \Delta], t \in \tau} |\hat{\beta}_p(s, t) - \beta_p(s, t)| \overset{p}{\to} 0$ as $n \to \infty$. 
The proof of this consistency, presented in Appendix B.4, demonstrates consistency of the estimators $\hat{\alpha}_{1k}(t)$ and $\hat{\alpha}_{2k}(t)$. This result, along with the representation of the regression parameter functions, is used to obtain the desired result. Condition [A1] is easily satisfied using common estimation approaches in the functional data literature (see Staniswalis and Lee (1998), Yao et al. (2005a), or Hall et al. (2006)). Condition [A2] can be satisfied by basis functions such as the B-splines among others (Efroymovich, 1999). Condition [A3] is satisfied by selecting a bounded $\lambda$-range when performing cross-validation. Condition [A4] holds under a number of common settings discussed in Appendix B.4. We note that these results are restricted to the class of regression parameter functions that can be expanded using $B_{s1,k}(s)$ and $B_{s2,k}(s)$ for $k \in \{1, ..., K\}$.

The functional response can then be estimated as,

$$
\hat{Y}_i(t) = \hat{\beta}_0(t) + \int_0^{\Delta_1} \hat{\beta}_1(s,t)X_{1i}(t-s)ds + \int_0^{\Delta_2} \hat{\beta}_2(s,t)X_{2i}(t-s)ds.
$$

(4.10)

Consistency of $\hat{Y}_i(t)$ follows from the following proposition and the proof is provided in Appendix B.4.

**Proposition 2** Let $X_1^i(t)$, $X_2^i(t)$ and $Y^*(t)$ be a set of predictors and corresponding response function with estimator $\hat{Y}^*(t)$ given by equation (4.10). If assumptions [A1] - [A4] hold and $X_1^i(\cdot)$ and $X_2^i(\cdot)$ are absolutely integrable, then the estimator $\hat{Y}^*(t)$ is consistent as $n \to \infty$ (i.e. $\sup_{t \in T} |\hat{Y}^*(t) - E[Y^*(t)|X_1^i(\cdot),X_2^i(\cdot)]| \overset{P}{\to} 0$ where $E[Y^*(t)|X_1^i(\cdot),X_2^i(\cdot)] = \beta_0(t) + \int_0^{\Delta_1} \beta_1(s,t)X_1^i(t-s)ds + \int_0^{\Delta_2} \beta_2(s,t)X_2^i(t-s)ds$).

An alternative approach to predict $Y_i(t)$ is to use an appropriate basis expansion of $X_{1i}(t)$ and $X_{2i}(t)$. To do this, assume each of the predictor functions are realizations of a square integrable process with continuous means, $\mu_1(t)$ and $\mu_2(t)$ and covariance functions, $G_{X_1,X_1}$ and $G_{X_2,X_2}$. Then for $p = 1, 2$, Mercer’s theorem yields the spectral decomposition of the covariance $G_{X,p,X,p}(s_1,s_2) = \sum_{l=1}^{\infty} \omega_{pl} \psi_{pl}(s_1)\psi_{pl}(s_2)$ in terms of non-negative eigenvalues $\omega_{pl} \geq \omega_{p2} \geq \cdots \geq 0$ and orthogonal eigenfunctions $\psi_{pl}(t)$ with $\int_T \psi_{pl}(t)\psi_{pl'}(t)dt = 1(l = l')$, where $1(l = l')$ is the indicator function which equals 1 when $l = l'$ and 0 otherwise (Bosq, 2000). Given that $E[X_{pl}(t)] = 0$ and the predictors are observed without noise, the above decomposition of the covariance implies that $X_{pl}(t)$ can be represented via the Karhunen-Loeve (KL) expansion as, $X_{pl}(t) = \sum_{l=1}^{\infty} \zeta_{pl,i}\psi_{pl}(t)$, where $\zeta_{pl,i} = \int_T X_{pl}(t)\psi_{pl}(t)dt$ are commonly called the functional principal component (FPC) scores and are uncorrelated random variables with zero mean and variance equal to $\omega_{pl}$. For practical and theoretical reasons, the KL expansion is often truncated (Yao et al., 2005a; Hall et al., 2006; Di et al., 2009); let $X_{11i}^{L1} = \sum_{l=1}^{L_1} \zeta_{1l,i}\psi_{1l}(t)$ and $X_{21i}^{L2} = \sum_{l=1}^{L_2} \zeta_{2l,i}\psi_{2l}(t)$ be the truncated KL expansions of $X_1(t)$ and $X_2(t)$ respectively. It follows that for $p = 1, 2$, $X_{pl}^{Lp}(t) \to X_{pl}(t)$ as $L_p \equiv L_p(n) \to \infty$ as $n \to \infty$ where the convergence is uniform.
in quadratic mean. Observe that substituting the above KL expansion into equation (4.2) yields the estimator of the response,
\[ \hat{Y}_{ij} = \hat{\beta}_0(t_j) + \sum_{k=1}^{K} \hat{\alpha}_{1k}(t_j) \int_0^\Delta \sum_{l=1}^{L_l} \hat{\zeta}_{l,i} \hat{\psi}_{1l}(t_j - s) B_{s1,k}(s) \, ds \]
\[ + \sum_{k=1}^{K} \hat{\alpha}_{2k}(t_j) \int_0^\Delta \sum_{l=1}^{L_l} \hat{\zeta}_{2l,i} \hat{\psi}_{2l}(t_j - s) B_{s2,k}(s) \, ds \]
\[ = \hat{\beta}_0(t_j) + \sum_{l=1}^{L_l} \hat{\zeta}_{l,i} \hat{P}_{1l}(t_j) + \sum_{l=1}^{L_l} \hat{\zeta}_{2l,i} \hat{P}_{2l}(t_j), \] (4.11)
where, \( \hat{\alpha}_{1k}(t_j), \hat{\alpha}_{2k}(t_j), \hat{\zeta}_{l,i}, \hat{\zeta}_{2l,i}, \hat{\psi}_{1l}(t_j - s), \) and \( \hat{\psi}_{2l}(t_j - s) \) are estimators of \( \alpha_{1k}(t_j), \alpha_{2k}(t_j), \zeta_{l,i}, \zeta_{2l,i}, \psi_{1l}(t_j - s), \) and \( \psi_{2l}(t_j - s) \) respectively, \( \hat{P}_{1l}(t_j) = \sum_{k=1}^{K} \hat{\alpha}_{1k}(t_j) \int_0^\Delta \hat{\psi}_{1l}(t_j - s) B_{s1,k}(s) \, ds \) and \( \hat{P}_{2l}(t_j) \) is defined similarly. Necessary conditions for pointwise convergence of this estimator are listed in Proposition 3 and the proof is provided in Appendix B.4.

**Proposition 3** Let \( X^*_1(t), X^*_2(t) \) and \( Y^*(t) \) be a set of predictors and corresponding response function with estimator \( \hat{Y}^*(t) \) given by equation (4.11). Assume \([A1]-[A4]\) hold and

[B1] the estimators for the functional eigenvectors and scores for the KL decomposition of \( X^*_1(t) \) and \( X^*_2(t) \) satisfy
\[ \sup_{t \in T} |\hat{\psi}_{1l}(t) - \psi_{1l}(t)| = O_p(n^{-\alpha}) \]
\[ |\hat{\zeta}^*_l - \zeta^*_l| \xrightarrow{p} 0 \quad \text{as} \quad n \to \infty \] (4.12)
for \( l \in \{1, \ldots, L_1\} \) and with similar conditions for \( \hat{\psi}_{2l} \) and \( \hat{\zeta}^*_l \).

[B2] Assume that \( \omega_{1l} = 0 \) for all \( l > L_1 \) for some \( L_1 \). Similarly assume \( \omega_{2l} = 0 \) for all \( l > L_2 \) for some \( L_2 \).

[B3] \( \beta_1(s,t) \) and \( \beta_2(s,t) \) are absolutely integrable functions.

Then, as \( n \to \infty, \left| \hat{Y}^*(t) - E[Y^*(t)|X^*_1(\cdot), X^*_2(\cdot)] \right| \xrightarrow{p} 0. \]

Note that conditions [B1] and [B2] are satisfied by estimators such as the ones introduced by Yao et al. (2005a) under some smoothness conditions on the predictors.

Selection of the smoothing parameters \( \lambda_1 \) and \( \lambda_2 \) is very important. To ensure that the regularization parameters are the same for all \( t_j \), these parameters can be selected using cross-validation (CV) to minimize an error based criterion, the Normalized Prediction Error,
\[ NPE_{(\lambda_1, \lambda_2)} = (1/n) \sum_{i=1}^{n} \left\{ \frac{1}{m} \sum_{j=1}^{m} \left| \hat{Y}_{ij} - Y_{ij} \right| \right\} / \left( \frac{1}{m} \sum_{j=1}^{m} |Y_{ij}| \right). \]

**Extensions**

Here we discuss the extension of the proposed method to more realistic settings where the predictors are possibly observed with noise and the sampling design for both the predictors and response is dense or sparse. First assume a dense sampling design for both the functional response and the predictors; that is \( m_{01}, m_{11}, \) and \( m_{21} \) are large. We note that the true predictor
functions $X_1(t)$ and $X_2(t)$ can be reconstructed with negligible error by obtaining smooth estimators $\hat{X}_1(t)$ and $\hat{X}_2(t)$ using local polynomial kernel smoothing (Zhang and Chen, 2007). In this setting, one can obtain predictions of the response by smoothing the observed data and using the empirical covariance and cross-covariance functions to obtain estimates of $G_{X_1X_1}(s_1,s_2)$, $G_{X_2X_2}(s_1,s_2)$, $G_{X_1X_2}(s_1,s_2)$, $G_{X_1Y}(s_1,s_2)$, and $G_{X_2Y}(s_1,s_2)$.

Another approach to handle predictors observed with noise is to use their KL decomposition. Recall from the previous section that given $E[X_{1i}(t)] = 0$ and $E[X_{2i}(t)] = 0$, the spectral decomposition of the predictor’s covariance functions implies that for $p = 1, 2$, $X_{pi}(t)$ can be represented via the KL expansion, $X_{pi}(t) = \sum_{l=1}^{\infty} \zeta_{pl,i} \psi_{pl}(t) + \gamma_{pi}(t)$, where $\gamma_{pi}(t)$ is an independent and identically distributed measurement error process with zero mean and variance $\sigma_{pi}^2$. Furthermore, the FPC scores are $\zeta_{pl,i} = \int_{\tau} X_{pi}(t) \psi_{pl}(t) dt$. The spectral decomposition of the smooth estimates of $G_{X_1X_1}(s_1,s_2)$ and $G_{X_2X_2}(s_1,s_2)$ yields the pairs of estimated eigenfunctions and eigenvalues $\{\hat{\psi}_{11}, \hat{\eta}_{11}\}$ and $\{\hat{\psi}_{21}, \hat{\eta}_{21}\}$ respectively with $\hat{\eta}_{11} > \hat{\eta}_{22} > ... > 0$ and $\hat{\eta}_{21} > \hat{\eta}_{22} > ... \geq 0$. Then one can obtain consistent estimators of the scores $\hat{\zeta}_{1l,i} = \int_{\tau} \hat{X}_{1i}(t) \hat{\psi}_{1l} dt$ and $\hat{\zeta}_{2l,i} = \int_{\tau} \hat{X}_{2i}(t) \hat{\psi}_{2l} dt$ needed for prediction (Hall et al., 2006; Zhang and Chen, 2007). Next, consider the case in which both predictor functions are observed on a sparse grid of points, but the response is observed for each subject on a sparse grid of points; that is, $m_{0i}$ is small for each $i$ but the set of all time points $\{t_{ij}\}_{ij}$ is dense in $[0,1]$. In this setting, one can obtain predictions of the response by smoothing the observed data and using the empirical covariance and cross-covariance functions to obtain estimates of $G_{X_1X_1}(s_1,s_2)$, $G_{X_2X_2}(s_1,s_2)$, and $G_{X_1X_2}(s_1,s_2)$. Since the response is observed on a sparse grid of points, bivariate kernel smoothing can be used to obtain $G_{X_1Y}(s_1,s_2)$, and $G_{X_2Y}(s_1,s_2)$ (Yao et al., 2005a; Wood, 2006). The FPC scores can be obtained as in the previous case.

As a third scenario, we consider when the response and predictor functions are observed on a sparse grid of points; that is, $m_{0i}, m_{1i},$ and $m_{2i}$ are small for each $i$ but the set of all time points $\{t_{ij}\}_{ij}$ is dense in $[0,1]$. To obtain smooth estimators of $G_{X_1X_1}(s_1,s_2)$, $G_{X_2X_2}(s_1,s_2)$, $G_{X_1X_2}(s_1,s_2)$, $G_{X_2X_2}(s_1,s_2)$, and $G_{X_1Y}(s_1,s_2)$, one can use common bivariate smoothing techniques, such as local linear smoothing or penalized spline smoothing (Yao et al., 2005a; Wood, 2006). Then the variance of the noise, $\sigma_1^2$ and $\sigma_2^2$, can be estimated based on the difference between the pointwise variance of the corresponding observed predictors and the estimated pointwise variance $\hat{G}_{X_1X_1}(s_1,s_2)$, $\hat{G}_{X_2X_2}(s_1,s_2)$ (Staniswalis and Lee, 1998). In this case, estimation of the scores via numerical integration is no longer reliable. In this setting, the conditional expectation formula should be employed, $\hat{\zeta}_{1l,i} = E[\hat{\zeta}_{1l,i}|X_1(t_{1i}),...,X_1(t_{m_{1i}})] = \hat{\eta}_{11} \hat{\psi}_{1l}(\hat{G}_{X_1X_1} + \hat{\sigma}_1^2 I_{m_{1i} \times m_{1i}})^{-1} X_{1i}$ (Yao et al., 2005a). Here $X_{1i}$ is a vector of length $m_{1i}$ containing the observed values of the first predictor function, $\hat{\psi}_{1l}^\top$ is a vector of length $m_{1i}$ with the $j$th entry equal to $\hat{\psi}_{1l}(t_{ij})$, $\hat{G}_{X_1X_1}$ is a $m_{1i} \times m_{1i}$-dimensional matrix with the $(j,j')$th entry
equal to $\tilde{G}_{X_1,X_2}(t_{ij},t_{ij'})$, and $I_{m_1 \times m_1}$ is the $m_1 \times m_1$ identity matrix. $\hat{\gamma}_{2l,i}$ can be estimated similarly. Yao et al. (2005a) showed that under the assumption that the predictors and their errors are jointly Gaussian, this equation yields the empirical best linear unbiased predictor of the scores.

Last, we consider the setting when the response is observed on a dense grid of points and the predictor functions are observed with noise for each subject on a sparse grid of points. In this case, the conditional expectation should be used to obtain estimates of the FPC scores and the predictor functions are observed with noise for each subject on a sparse grid of points. The estimators described in this section can be used to obtain predictions under this pointwise estimation framework. In the next section, we propose an alternative approach to estimating the lagFLM model.

### 4.2.2 GB Model Representation

In this section we propose the GB method which takes a global smoothing approach to estimate the regression coefficient functions of the lagFLM model. To do this, we expand the time varying parameters in (4.3) by taking $\beta_0(t) = \sum_{l=1}^{L_0} B_{0l,t}(t)b_{0l}$, $\alpha_{1k}(t) = \sum_{l=1}^{L_1} B_{1l,t}(t)b_{1kl}$, and $\alpha_{2k}(t) = \sum_{l=1}^{L_2} B_{2l,t}(t)b_{2kl}$ where $L_0$, $L_1$ and $L_2$ are large enough constants; $\{B_{0l,t}(t)\}$, $\{B_{1l,t}(t)\}$ and $\{B_{2l,t}(t)\}$ are one dimensional basis functions; and $b_{0l}$, $b_{1kl}$, and $b_{2kl}$ are unknown basis coefficients. This choice of representation allows one to express the coefficient function as $\beta_1(s,t) = \sum_{k=1}^{K} \sum_{l=1}^{L_1} B_{s1,k}(s)B_{1l,t}(t)b_{1kl}$ which uses a bivariate basis expansion obtained from a tensor product of two univariate bases. Nevertheless, the GB method can accommodate for other bivariate basis functions, $\{B_{1l',t}(s,t)\}_{l'=1}^{L_1}$, such as the bivariate thin plate regression splines (TPRS) used in the simulation experiments and data analysis. For simplicity, consider the latter approach and represent the bivariate effects as a linear combination of bivariate basis functions: $\beta_1(s,t) = \sum_{l=1}^{L_1} B_{1l,t}(s,t)b_{1l}$ for known functions $B_{1l,t}(s,t)$ and unknown coefficients $b_{1l}$. Similarly, we take $\beta_2(s,t) = \sum_{l=1}^{L_2} B_{2l,t}(s,t)b_{2l}$ for known functions $B_{2l,t}(s,t)$ and unknown coefficients $b_{2l}$. With an abuse of notation we redefined $L_1$ and $L_2$ to be the number of bivariate basis.

Similar to Section 4.2.1, the conditional mean, $f_i(t_j)$, can be represented as,

$$f_i(t_j) = \sum_{l=1}^{L_0} B_{0l,t}(t_i)b_{0l} + \sum_{l=1}^{L_1} \bar{X}_{1l,i}(t_j)b_{1l} + \sum_{l=1}^{L_2} \bar{X}_{2l,i}(t_j)b_{2l}$$

(4.13)

where $\bar{X}_{1l,i}(t_j) = \int_0^{t_j} X_{1l,i}(t_j - s)B_{1l,t}(s,t_j) \, ds$ and $\bar{X}_{2l,i}(t_j)$ is defined analogously. This representation is inspired by Ivanescu et al. (2014) which presents a penalized function on function regression framework that uses entire range of predictor functions. Equation (4.13) looks similar to equation (4.3); the main difference is that the unknown coefficients, $b_{0l}$, $b_{1l}$, and $b_{2l}$ are not time varying. This modeling approach ensures smoothness of the regression parameter.
functions $\beta_1(s, t)$ and $\beta_2(s, t)$ over both $s$ and $t$.

The coefficient parameter vectors, $b_0 = [b_{01}, \ldots, b_{0L_0}]^\top$, $b_1 = [b_{11}, \ldots, b_{1L_1}]^\top$, and $b_2 = [b_{21}, \ldots, b_{2L_2}]^\top$ are estimated by minimizing the following penalized criterion,

$$\sum_i \{Y_i(t_j) - f_i(t_j)\}^2 + \lambda_0 P_0(b_0) + \lambda_1 P_1(b_1) + \lambda_2 P_2(b_2), \quad (4.14)$$

where $\lambda_0$, $\lambda_1$, and $\lambda_2$ are the smoothing parameters and the penalty terms are $P_1$ and $P_2$ (Wood, 2011; Ivanescu et al., 2014). In particular, the integral of the square of the second derivative is a common measure for smoothness, in which case $P_0(b_0) = \int_\tau \{\beta_p^0(t)\}^2 \, dt = b_0^\top S_0 b_0$ and $P_p(b_p) = \int_\tau \int_0^{\Delta p} \left[ \{\partial^2 \beta_p(s, t)/\partial s^2\}^2 + \{\partial^2 \beta_p(s, t)/\partial t^2\}^2 \right] \, ds \, dt = b_p^\top S_p b_p$ where $S_0$ and $S_p$ are penalty matrices specified by the choice of basis functions (Wood, 2006).

Estimation can be performed by selecting optimal regularization parameters via generalized cross validation (GCV) as done for the $PW$ method. However, empirical evidence suggests that GCV leads to undersmoothing (Reiss and Todd Ogden, 2009; Wood, 2011). Instead, the problem can be posed as a mixed model of the form $X \cdot b_f + Z r$, where $b_f$ contains the fixed effect parameters and $r$ is a random effect vector (Wood, 2006, 2011). In this context, $\lambda_0$, $\lambda_1$, and $\lambda_2$ represent variance components of the random effect covariance structure. This framework, allows us to use a restricted maximum likelihood (REML) estimation procedure (Reiss and Todd Ogden, 2009; Wood, 2006, 2011) to solve this problem. This approach is also more robust to non-independent error assumptions (Krivobokova and Kauermann, 2007; Ivanescu et al., 2014).

For the dense scenario, we can directly compute $\hat{X}_{pl,i}$ and we use REML to obtain estimates of $b_{pli}$. Under the REML framework, we know that under some regularity assumptions $|\hat{b}_{pl} - b_{pl}| \xrightarrow{P} 0$ (Cressie and Lahiri, 1993). Therefore, if $B_{1,i}(s, t)$ is uniformly bounded (similar to assumption [A2] from Proposition 1) we have that sup$_{t,s}|\hat{\beta}_1(s, t) - \beta_1(s, t)| \xrightarrow{P} 0$ and similar for $\beta_2(s, t)$.

Predictions of the functional response can be obtained using the same methodology presented for the $PW$ method. In particular, we let

$$\hat{Y}_i(t) = \hat{\beta}_0(t) + \int_0^{\Delta 1} \hat{\beta}_1(s, t) X_{1i}(t - s) \, ds + \int_0^{\Delta 2} \hat{\beta}_2(s, t) X_{2i}(t - s) \, ds.$$  

This is the same estimator that was shown to be consistent in proposition 2.

**Extensions**

Here we discuss the extension of the proposed $GB$ method to more realistic settings where the predictors are possibly observed with noise and the sampling design for both the predictors and response is dense or sparse. First, assume a dense sampling design for both the functional response and the predictors; that is $m_{0i}$, $m_{1i}$, and $m_{2i}$ are large. We recall that the true predictor
functions $X_1(t)$ and $X_2(t)$ can be reconstructed with negligible error by obtaining smooth estimators $\hat{X}_1(t)$ and $\hat{X}_2(t)$ using local polynomial kernel smoothing (Zhang and Chen, 2007). Therefore, we simply can fit the model with the smooth trajectories, $\hat{X}_1(t)$ and $\hat{X}_2(t)$, using the proposed method. This same approach can be taken in the case when the response is observed for each subject on a sparse grid of points, but the predictor functions are observed with noise for each subject on a dense grid of points.

Next, assume a sparse sampling design for both the functional response and the predictors; that is $m_0, m_1,$ and $m_2$ are possibly very small, but the set of all time points $\{t_{ij}\}_{i,j}$ is dense in $[0, 1]$. In order to get dense estimates of the predictor functions, which is required for the GB approach, the observed data must first be smoothed. We assume each of the predictor functions are realizations of a square integrable process with continuous mean and covariance function, $G_{X, X}$. Therefore, following the same argument as in Section 4.2.1, common FPCA techniques can be applied to obtain the smooth trajectories (Di et al., 2009; Goldsmith et al., 2012; Staniswalis and Lee, 1998; Yao et al., 2005b). The reconstructed predictor trajectories are then dense, so they can be used to fit the model. This approach can also be employed in the case when the response is observed on a dense grid of points and the predictor functions are observed with noise for each subject on a sparse grid of points.

### 4.3 Simulation Studies

This section presents results from an extensive simulation study used to evaluate the performance of the two proposed methods, PW and GB, in terms of prediction performance. We consider two settings for data generated as in model (4.1); (I) the response and functional predictors are observed densely so that $m_0, m_1,$ and $m_2$ are large and the predictors are observed without measurement error; and (II) the functional predictors and response are observed on a sparse grid of points so that $m_0, m_1,$ and $m_2$ are small and the predictors are observed with measurement error.

We first generate the response and predictor functions on a dense and equally spaced grid of points, $t_{ij} \in T = \{j/99 : j = 0, 1, 2, \ldots, 99\}$. In setting (I) we generate observed data, $\{t_{ij}, Y_{ij}, W_{1ij}, W_{2ij}\}_{j}$, where $Y_{ij}$ is the response function and $W_{1ij} = X_{1i}(t_{ij})$ and $W_{2ij} = X_{2i}(t_{ij})$ are the predictor functions measured without error. We define the predictors, $X_{1i}(t) = \sum_{k=1}^{2} u_{ik1}\sin(k\pi t) + u_{ik2}\cos(k\pi t)$ and $X_{2i}(t) = \sum_{k=1}^{3} v_{ik}\cos(2k\pi t)$ for $u_{ik1}, u_{ik2},$ and $v_{ik}$ $\sim N(0, 1/k^4)$. For our simulations we use regression parameter functions $\beta_0(t) = e^{-(t-5)^2}$, $\beta_1(s, t) = \cos(2\pi t)\cos(\pi s)$ and $\beta_2(s, t) = \cos(\pi t)\cos(\pi s)$ and $\Delta = 0.4$ such that $s \in \{j/99 : j = 0, 1, 2, \ldots, [100\Delta]\}$.

The response for subject $i \in 1, \ldots, n$ is generated from model (4.1) via numerical integration for each $t_{ij} \in T$. The measurement error in model (4.1) is taken to be $\epsilon_i(t) = \epsilon_{a,i}(t) + \epsilon_{g,i}(t)$ where
\(\epsilon_{g,i}(t)\) is a zero-mean iid normal random variable with variance \(\sigma^2/2\), and \(\epsilon_{a,i}(t)\) is a zero-mean AR(1) process with covariance structure given by \(\Sigma(t_j, t_{j'}) = (\sigma^2/2)\rho^{\mid j-j'\mid}\) with \(\rho = 0.5\). Note that for a fixed \(t\), the variance of \(\epsilon_i(t)\) is equal to \(\sigma^2\). We simulate data under a framework similar to Ivanescu et al. (2014) and Mclean et al. (2014), by using a version of the empirical signal to noise ratio (eSNR\(_\epsilon\)), which we define as 

\[
esSNR_{\epsilon} = \left(\frac{\sum_{i,j} \{E(Y_{ij}) - \sum_i E(Y_{ij})/n\}^2}{\sigma^2 \epsilon(n-1)m}\right).
\]

The value of \(\sigma^2\) is chosen by specifying the eSNR\(_\epsilon\) for each Monte Carlo (MC) replication. In this section, we present results for eSNR\(_\epsilon\) = 5. It was observed that the predictive error decreases (increases) as eSNR\(_\epsilon\) is increased (decreased). Results for eSNR\(_\epsilon\) = 1 as well as for responses generated with a white noise error process are presented in the appendix.

In setting (II) the response and predictors are first generated as in setting (I). Then to generate the sparse data we evaluate the functions at points \(r_{ik} \in \mathcal{R}_i\), for a randomly sampled subset \(\mathcal{R}_i \subset \mathcal{T}\) of size \(m_i\). To determine the number of observations for each subject we take \(m_i \sim \text{Uniform}\) on \(\{10, 11, \ldots, 15\}\). To generate the predictors that are observed with error we take \(W_{1ik} = X_{1i}(r_{ij}) + \epsilon_{1ik}\) and \(W_{2ik} = X_{2i}(r_{ij}) + \epsilon_{2ik}\) where \(\epsilon_{1ik} \sim N(0, 0.1)\) and \(\epsilon_{2ik} \sim N(0, 0.05)\). The latter error variances were selected to correspond to a signal to noise ratio of five.

The main simulation factors of interest are the number of observations and the lag parameter used for fitting the model. We present results for \(n \in \{20, 50, 100\}\). To understand how the model performs when the lag parameter is mis-specified we let \(\Delta_F\) be the lag parameter used for fitting the model and present results for \(\Delta_F \in \{0.2, 0.4, 0.6\}\). It was found that both estimation methods have poor predictive accuracy when \(\Delta_F < \Delta\) as this corresponds to fitting the model with missing data not at random. This is equivalent to setting \(\beta_1(s,t) = 0\) and \(\beta_2(s,t) = 0\) for \(s \in [\Delta_F, \Delta]\) which we found to increase the bias in predictions. Accounting for this case is a possible area of future work, but is not of immediate interest as one can set \(\Delta_F\) to be larger than necessary.

To visualize the data analyzed in these experiments, some figures of the simulated data and corresponding predictions are presented in appendix B.1. Figure B.1 presents the observed predictor and response functions for simulation settings (I) and (II) from a randomly chosen MC replications with \(n = 100\). Figure B.2 presents the corresponding smooth predictor trajectories. In figure B.3, we see that the predicted responses using both approaches for simulation settings (I) and (II) are similar.

To fit the model, the predictors are first standardized to have a zero mean function and pointwise variance equal to 1. For the model using the PW method, we consider the basis set \(\{B_{s1,k}\}\) is taken to be the set of \(K = 10\) B-spline basis functions of degree 4 with 6 interior knots over \([0, \Delta_F]\). This value of \(K\) was selected empirically. We use one tuning parameter \(\lambda\), that is selected using a grid search over \(I_\lambda = [0.0001, 0.01]\) and GCV. The range was determined
empirically and using a larger range as well as two tuning parameters provided similar results. To obtain the smooth covariance and cross covariance functions, \( \hat{G}_{X_1X_1}(s_1,s_2), \hat{G}_{X_1X_2}(s_1,s_2), \hat{G}_{X_2X_1}(s_1,s_2), \hat{G}_{X_2X_2}(s_1,s_2), \) and \( \hat{G}_{X_1Y}(s_1,s_2), \) we use bivariate smoothing, via tensor products of univariate bases. The positive semi-definiteness of the covariance functions is ensured by truncating the eigenvalues to be positive. The results are based on using the tensor product of 20 univariate cubic regression splines with 15 equally spaced interior knots.

For the GB method, the functions \( B_{p,k}(s,t) \) are chosen to be 30 thin plate spline basis functions with 30 knots equally spaced. The penalty for equation (4.14) is taken to be \( P_l(b_l) = b_l^\top S_l b_l \) where \( S_l \) is chosen to be the thin plate spline penalty (Wood, 2006). In setting (II), the reconstructed trajectories are estimated using FPCA with 20 basis functions to smooth the covariance and cross covariance functions and the percent variance explained set to 99%.

To obtain the necessary smooth estimates we use \code{fpca.sc} function of the package \textit{refund} (Di et al., 2009; Goldsmith et al., 2012; Staniswalis and Lee, 1998; Yao et al., 2005b). To obtain predictions we use the \code{gam} function of the \textit{mgcv} package (Wood, 2006, 2011).

The computational cost and empirical performance of the PW method depends directly on the range of values \( I_\lambda \) used to search for the optimal regularization parameter. Since the estimation of regression parameters for the PW method uses only the smooth covariance and cross-covariance estimates at each time \( t_{ij} \), it is more computationally efficient than the GB method which uses large matrices that grow linearly with \( \sum_{i=1}^n m_i \). See appendix B.6.1 and B.6.2 for algorithms that outline the implementation details.

### 4.3.1 Simulation Results

To evaluate model performance we present the relative integrated mean square error (rMSE) for the \( \hat{\beta}_1(s,t) \) and \( \hat{\beta}_2(s,t) \) functions. For \( p = \{1,2\} \),

\[
\text{rMSE}(\hat{\beta}_p) = \frac{\int_0^T \int_\Delta \{ \hat{\beta}_p(s,t) - \beta_p(s,t) \}^2 \, ds \, dt}{\int_0^T \int_\Delta \{ \beta_p(s,t) \}^2 \, ds \, dt}.
\]

The rMSE for \( \hat{\beta}_0(t) \) is defined similarly. The rMSE for \( Y(t) \) is given by, \( \text{rMSE}(Y) = \frac{\sum_i \int_\Delta \{ \hat{Y}_i(t) - f_i(t) \}^2 \, dt}{\sum_i \int_\Delta \{ f_i(t) \}^2 \, dt} \). We report the mean of the rMSE(\( \beta_p \)) and rMSE(\( Y \)) values over 500 Monte Carlo (MC) replications. Since our main objective is prediction, the standard error is reported for the rMSE(\( Y \)) as well. We compare the resulting prediction errors by evaluating improvement of quantity ‘A’ over ‘B’, defined to be \( [(B-A)/B] \cdot 100\% \).

First we describe and compare the results for the two estimation methods under the dense setting. Tables 4.1 and 4.2 report the simulation results for the PW and the GB methods respectively under setting (I). On average, the GB method yields an improvement of 92% over the PW method in this setting. For the PW method we observe that the case when \( \Delta_F = \Delta \) has smaller prediction errors than the case when \( \Delta_F \geq \Delta \); observe that on average the former case
yields a 15% improvement over the latter. For the GB method, on average having $\Delta_F = \Delta$ yields a 20% improvement over having $\Delta_F \geq \Delta$. As expected, the prediction error decreases in both cases as the sample size increases. For the PW method, $n = 50$ yields an average improvement of 47% over the case of $n = 20$, and for $n = 100$ we observe an average improvement of 36% over the case when $n = 50$. Using the GB method with $n = 50$ yields an average improvement of 56% over the case when $n = 20$; however, for $n = 100$ we observe an average improvement of 48% over the case when $n = 50$.

Next, we compare performance under the dense and sparse settings for each of the methods. Tables 4.1 and 4.3 report the simulation results for the PW method under settings (I) and (II) respectively. From these tables, observe that for the PW method the prediction errors in setting (I) are on average 70% smaller than in setting (II). Similarly, tables 4.2 and 4.4 show that for the GB method the prediction errors in setting (I) are on average 95% smaller than in setting (II).

Last we describe and compare the results for the two estimation methods under the sparse setting. Tables 4.3 and 4.4 report the simulation results for the PW and the GB methods respectively under the sparse setting (II). On average, the GB method yields an improvement of 55% over the PW method. For PW we observe that the case when $\Delta_F = \Delta$ has smaller prediction errors than the case when $\Delta_F \geq \Delta$; observe that on average the former case yields a 12% improvement over the latter. For GB on average having $\Delta_F = \Delta$ yields a 1% improvement over having $\Delta_F \geq \Delta$. As expected, the prediction error decreases in both cases as the sample size increases. For PW , $n = 50$ yields an average improvement of 47% over the case when $n = 20$, and for $n = 100$ we observe an average improvement of 30% over the case when $n = 50$. For GB , $n = 50$ yields an average improvement of 48% over the case when $n = 20$, and when $n = 100$ we observe an average improvement of 17% over the case when $n = 50$.

Overall the GB method has smaller predictive errors for both dense and sparse settings. Additionally, we also observe performance is better when $\Delta_F = \Delta$ compared to when $\Delta_F > \Delta$. We also see improvement in performance as the number of observations increase.

Table 4.1: Prediction Error of PW Method - Setting (I), Average Error Variance = 0.012

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\Delta$</th>
<th>$\Delta_F$</th>
<th>$rMSE(\beta_0)$</th>
<th>$rMSE(\beta_1)$</th>
<th>$rMSE(\beta_2)$</th>
<th>$rMSE(Y)$ ($SE[rMSE(Y)]$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.4</td>
<td>2.9e-05</td>
<td>1</td>
<td>2.1</td>
<td>5.1e-03 ( 5.9e-03 )</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.6</td>
<td>2.6e-05</td>
<td>2.1</td>
<td>2.3</td>
<td>6.5e-03 ( 0.011 )</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.4</td>
<td>1.2e-05</td>
<td>0.96</td>
<td>2.2</td>
<td>2.9e-03 ( 4.6e-03 )</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.6</td>
<td>1.1e-05</td>
<td>0.73</td>
<td>1.4</td>
<td>3.2e-03 ( 2.8e-03 )</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.4</td>
<td>5.8e-06</td>
<td>0.23</td>
<td>1.1</td>
<td>1.8e-03 ( 1.4e-03 )</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.6</td>
<td>5.4e-06</td>
<td>0.54</td>
<td>1.2</td>
<td>2.1e-03 ( 2e-03 )</td>
</tr>
</tbody>
</table>
4.4 Prediction of Magnetization Transfer Ratio within Lesions of MS Patients

Recall that the data consist of MTR maps, T1w, and FLAIR images obtained for 53 MS patients who were imaged as part of a natural history study at the National Institute of Neurological Disorders and Stroke. The patients were imaged between 9 - 38 times over a period of 5.5 years. Our objective is to evaluate the performance of the lagFLM model using longitudinal voxel-level measurements of the T1w and FLAIR images over the previous six months of time to obtain
predictions of MTR at a specified time within lesions. The window is chosen to be six months because it is known that the progression of tissue damage at a specified time within a lesion is only prominent within the previous six months (Van Den Elskamp et al., 2008). We evaluate how the lagFLM model performs with each of the predictors compared with two cross-sectional models that do not take temporal dependence of the data into account. Section 4.4.1 discusses the pre-processing of the data and Section 4.4.2 presents implementation details and results for this analysis.

4.4.1 Data Processing

We apply our methods to a data set consisting of images taken on 53 patients as part of a natural history study at the National Institute of Neurological Disorders and Stroke. The patients were imaged between 9 - 38 times over a period of 5.5 years. During each visit each of the following was acquired: T1-weighted spoiled gradient echo (Volumetric-T1), T2-weighted FLuid Attenuated Inversion Recovery (FLAIR), and two T1-weighted sequences, with (MT-ON) and without (MT-OFF) an off-resonance magnetization transfer pre-pulse (for use in MTR calculation). We applied the N4 inhomogeneity correction algorithm (Tustison et al., 2010) to the Volumetric-T1 and FLAIR volumes, and extracerebral voxels are removed using the SPECTRE skull-stripping algorithm (Carass et al., 2011) on the Volumetric-T1 image and the resulting mask was applied to all images. To spatially register the images, for each subject, we rigidly align the corrected MT-OFF, Volumetric-T1 and FLAIR volumes to the MNI152 1.0 mm nonlinear template, using a two-step registration technique where a second alignment is done without the skull. The transform from the MT-OFF image was then used to transform the MTR map into the common space.

Since the T1w and FLAIR images are measured without meaningful units, they must be normalized so that repeated scans are measured on the same scale. To accomplish this, each observed image is normalized using the methods of Shinohara et al. (2011b). At every time point the means and variances of each of the T1w and FLAIR images are computed within a region of normal appearing white matter (NAWM). Then each image is normalized by subtracting the corresponding mean and dividing by the corresponding standard deviation.

To temporally register the voxel trajectories, we align each voxel to the time point when it first is identified as belonging to a new or enlarging lesion. This time of incidence for each voxel is identified using the Subtraction-Based Logistic Inference for Modeling and Estimation (SuBLIME) procedure (Sweeney et al., 2013a). For each voxel, we obtain the time marker that corresponds to the time when the voxel is first identified as being part of a lesion (‘time zero’). We then measure all following time points as the time difference in weeks from this marker. This alignment is appropriate because our goal is to predict MTR in the time window after the
4.4.2 Model Implementation and Results

After normalization and spatio-temporal registration, we let $Y_i(t_{ij}, v_{ijk})$ be the MTR in the brain of subject $i$ at the time in weeks $t_{ij} \in [0, 190]$. We consider this measurement for voxels $v_{ijk} \in \mathcal{L}_i$, the collection of voxels measured at all times $t_{ij} \in [0, 190]$ for patient $i$. Similarly, we let the predictors be $W_{1i}(t_{ij}, v_{ijk})$ and $W_{2i}(t_{ij}, v_{ijk})$ the normalized T1w and FLAIR intensities respectively, of voxel $v_{ijk}$ for subject $i$ at time $t_{ij}$ which are assumed to be observed with noise. We consider these two predictors as they are commonly acquired images as part of clinical MRI studies.

Let $\Delta_F = 26$ weeks, the time between study visits that is believed to be of scientific importance (approximately six months). It is reasonable to assume that subjects are independent. Our modeling approach relies on a working independence assumption across voxels of MTR conditionally on the observed conventional imaging. Hence, we abridge our notation by considering response and predictor values of the form $Y_i(t_{ij}), W_{1i}(t_{ij})$ and $W_{2i}(t_{ij})$ where $i$ is the index for the voxels over all subjects, and $t_{ij}$ are the scan times measured in weeks with respect to the incidence time for the $i$th voxel. We analyze the data for $t_{ij} \in [0, 190]$ to ensure that there are sufficiently dense observations over the time domain.

Figure 4.1 presents images for a lesion of one patient. From top to bottom the T1w, FLAIR, and MTR images are displayed for an MS patient obtained over time in weeks from left to right (indicated below in white text). Time zero corresponds to lesion incidence. The lesion is the hypointense region in the center of the image on the T1w and MTR images and hyperintense on the FLAIR. Figure 4.2 (a), 4.2 (b), and 4.3 (a) present the corresponding T1w, FLAIR, and MTR voxel-level trajectories for this image.

In order to determine the benefit of the lagFLM model for predicting MTR, we also compare its performance against a cross-sectional model (CS),

$$ Y_i(t_{ij}) = \beta_0 + \beta_1 W_{1i}(t_{ij}) + \beta_2 W_{2i}(t_{ij}) + \epsilon_i(t_{ij}). \quad (4.15) $$

Additionally, we compare the performance against a cross-sectional non-linear model (CSNL),

$$ Y_i(t_{ij}) = \beta_0 + g_1[W_{1i}(t_{ij})] + g_2[W_{2i}(t_{ij})] + \epsilon_i(t_{ij}), \quad (4.16) $$

where $g_1(\cdot)$ and $g_2(\cdot)$ are unknown non-linear functions. These models do not account for any temporal dependencies that may be present in the data. To fit the cross-sectional models, the predictors are standardized to have mean zero and standard deviation of one. Predictions for the CSNL model are obtained using the gam function of the mgcv package in R with 10 thin plate basis functions with 10 equally spaced knots.

To fit the lagFLM models, the predictors are first standardized to have a zero mean function...
and pointwise variance equal to 1. The analysis is performed on the logit transform of the MTR trajectories since the response values lie in the range \([0, 1]\). The same analysis is also conducted on the raw data which yields comparable results (omitted). There are some zero values in the response but no values of 1, and most of the MTR values are concentrated between 0.2 and 0.5. Hence, in order to avoid division by zero we shift the data by 0.01 when performing the logit transformation. That is, for our analysis we define, \( \text{logit}(y) = \log \left( \frac{y + 0.01}{1 - (y + 0.01)} \right) \), and \( \text{logit}^{-1}(\alpha) = \frac{e^\alpha}{e^\alpha + 1} - 0.01 \). When using the logit option in the implementation, the response data is transformed before fitting the model. Then, predictions are transformed back to the probabilities using the \( \text{logit}^{-1} \) function.

For the PW method, the basis set \( \{B_{s1,k}\} \) and \( \{B_{s2,k}\} \) are taken to be the set of \( K = 10 \) B-spline basis functions of degree 4 with 6 interior knots over \([0, 26]\). The value of \( K \) was selected empirically and results were similar to \( K = 20 \). We use one tuning parameter \( \lambda \), that is selected using GCV over \([0.1, 100]\). To obtain the smooth covariance and cross-covariance functions, \( \hat{G}_{X_1X_1}(s_1, s_2) \), \( \hat{G}_{X_2X_2}(s_1, s_2) \), \( \hat{G}_{X_1X_2}(s_1, s_2) \), \( \hat{G}_{X_1Y}(s_1, s_2) \), and \( \hat{G}_{X_2Y}(s_1, s_2) \) we use bivariate smoothing, via tensor products of univariate bases. The positive semi-definiteness of the covariance and cross-covariance functions is ensured by truncating the eigenvalues to be positive. The results are based on using the tensor product of 20 univariate cubic regression splines with 15 equally spaced interior knots.

For the GB method, the functions \( B_{p,k}(s, t) \) are chosen to be 30 thin plate spline basis functions with 30 knots equally spaced. The penalty for equation (4.14) is taken to be \( P_l(b_l) = b_l^\top S_l b_l \) where \( S_l \) is obtained by using the thin plate spline penalty (Wood, 2006). The reconstructed trajectories are estimated for \( t_{ij} \in [0, 190] \) using FPCA with 20 basis functions to smooth the covariance and cross-covariance functions and the percent variance explained set to 99%. All analysis is conducted using \( \mathcal{R} \) version 2.15.1. To obtain the necessary smooth estimates we use the \textit{fpca.sc} function of the package \textit{refund} (Di et al., 2009; Goldsmith et al., 2012; Staniswalis and Lee, 1998; Yao et al., 2005b). To obtain predictions we use the \textit{bam} function of the \textit{mgcv} package (Wood, 2006, 2011).

Using this smoothing procedure we are able to capture the natural variation in the predictor functions. Figures 4.2 (a) and (b) display the observed and voxel-level predictor trajectories of the T1w and FLAIR images respectively for one patient in a subset of voxels and the respective smoothed trajectories are presented in Figure 4.2 (c) and (d). The voxel-level trajectories in these figures correspond to the voxels in the lesion displayed on the true images in Figure 4.1.

Since true values of the response are not available for the analysis and there is only a limited number of data available, we perform a 10-fold cross validation over subjects. The subjects are first randomly sorted and then split into groups by subject. The first three groups consist of six subjects and the other seven groups consist of five subjects.
We evaluate prediction by comparing the integrated mean squared error (MSE) of the models. We also compare this prediction error with the empirical measurement error of the MTR voxel-level trajectories. To do this, for each cross-validation step we compute \( MSE/s_0^2 \), such that \( s_0^2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} \left( Y_i(t_{ij}) - \tilde{Y}_i(t_{ij}) \right)^2}{\sum_{i=1}^{n} m_i} \) where \( m_i \) is the number of time samples per curve, and \( n \) is the total number of observed curves (including voxels and subjects), and \( \tilde{Y}_i(t_{ij}) \) are the smoothed response trajectories without noise. A ratio of 1 indicates that the prediction is as good as smoothing the observed data. The smooth trajectories are obtained by using the \( fpca.sc \) function in the \textit{refund} package in \texttt{R} with 20 basis functions to smooth the covariance and 99% variance explained to obtain the number of eigenfunctions used to reconstruct the data.

Figure 4.3 (a) and (b) display the observed and smoothed MTR voxel-level trajectories respectively for the voxels in the lesion of Figure 4.1. For these same voxel-level trajectories, Figure 4.3 (c) and (d) present the predicted response trajectories from using the \( PW \) and \( GB \) methods. We see that the predictions obtained using our model which only uses the values of T1w and FLAIR voxel intensities are similar to the estimated trajectories obtained by smoothing the observed data.

Table 4.5 displays the results for all the different approaches using different combinations of the two predictors and methodologies. The \textit{lag}FLM model consistently has higher predictive accuracy than the corresponding cross-sectional models. Both estimation approaches for the \textit{lag}FLM model perform comparably. In the case of two predictors, \textit{lag}FLM model yields a 64% improvement over the \textit{CS} model and a 46% improvement over the \textit{CSNL} model. This indicates that accounting for temporal dependence over the previous six months is important in predicting MTR.

Table 4.5: Prediction error from 10-fold cross validation on MTR data with two predictors. Integrated mean squared error (MSE) times 10^3 along with the average \( MSE/s_0^2 \), where \( s_0^2 \) is the estimated measurement error in the observed MTR voxel trajectories, for all models and estimation approaches presented in the paper.

<table>
<thead>
<tr>
<th>Model</th>
<th>Predictors</th>
<th>( MSE \ [SE] \cdot (10^3) )</th>
<th>( MSE/s_0^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td>T1w + FLAIR</td>
<td>14.0 [3.6]</td>
<td>4.5</td>
</tr>
<tr>
<td>CSNL</td>
<td>T1w + FLAIR</td>
<td>9.3 [2.9]</td>
<td>3</td>
</tr>
<tr>
<td>\textit{lag}FLM ( GB )</td>
<td>T1w + FLAIR</td>
<td>5.0 [1.8]</td>
<td>1.6</td>
</tr>
<tr>
<td>\textit{lag}FLM ( PW )</td>
<td>T1w + FLAIR</td>
<td>5.0 [1.8]</td>
<td>1.6</td>
</tr>
</tbody>
</table>

We also compare performance of the model when using both predictors compared to using only one in Table B.25 in Appendix B.7. It may be of future interest to explore the model with
interactions as well as other commonly acquired predictors such as proton attenuation volumes or T2w images. This analysis can provide insight for researchers regarding which commonly acquired images can be used to predict MTR and analyze tissue damage. Observe that using both predictors in the lagFLM model yields a 9% improvement over the model with only one predictor. However, when using the cross-sectional models, this relationship is not maintained; each of the two-predictor models do not perform as well as the corresponding model with only the T1w intensities used as a predictor. The models with only FLAIR intensities used as a predictor consistently does worse than all of the corresponding cross-sectional two-predictor models. These results confirm the qualitative assessment that the T1w image on its own contains detailed information about tissue damage. Furthermore, adding the FLAIR intensities to any of the cross-sectional models does not improve predictive power.

Figure 4.1: From top to bottom, axial slices of an MS lesion in the FLAIR (A), T1w (B), and MTR (C) sequences. The MS patient is imaged over time in weeks from left to right. Time zero corresponds to lesion incidence and the lesion on each sequence is denoted at time zero with a red arrow. On the MTR and T1w images the lesion is the hypointense region; on the FLAIR the lesion is hyper intense.
Figure 4.2: (a) T1w voxel intensities for a subset of the lesion voxels in one patient; (b) corresponding smoothed T1w voxel intensities using 20 basis functions to smooth all of the data; (c) corresponding FLAIR voxel intensities; (d) corresponding smoothed FLAIR voxel intensities using 20 basis functions to smooth all of the data.

Figure 4.3: (a) MTR voxel intensities for a subset of the lesion voxels in one patient; (b) smoothed MTR voxel intensities using 20 basis functions to smooth all of the data; (c) predicted MTR voxel intensities using the GB method; (d) predicted MTR voxel intensities using the PW method.
Chapter 5

Conclusion

In Chapter 2 we present a two-sample testing procedure to determine if certain diffusion tensor image modalities arise from different underlying processes. When dealing with functional data, which is infinite dimensional, it is important to use data reduction techniques that take advantage of the functional nature of the data. We propose the use of marginal FPCA to represent the curves using the eigenbasis of the marginal covariance of an appropriate mixture process. This reduces the dimension of the testing problem and allows us to apply well known lower-dimensional procedures. The proposed testing approach combines marginal FPCA and classical univariate procedures (e.g. Anderson-Darling test), scales well to larger samples sizes, and can be easily extended to test the null hypothesis that multiple (as in more than two) groups of curves have identical distribution. We found that the KS test has similar attributes but was not as powerful for detecting changes in the higher order moments of the coefficient distributions. Furthermore, we have shown that the proposed FAD test outperforms the CVM test of Hall and Van Keilegom (2007) for smaller sample sizes.

In Chapter 3, we propose a model that uses a historical covariate to predict lesion enhancement using contrast-free imaging. The gain in predictive performance that we observed is unprecedented, despite the simplicity and strong parametric assumptions of the model. Since using the historical covariate yields such an improvement. Requiring estimation of such a historical covariate imposes that prior scans must have been collected. Therefore, the proposed model is not applicable for patients who are only scanned at one visit. This would not be a limitation in two-arm placebo-controlled trials where only post-randomization scans are relevant, as long as there is a baseline scan.

The proposed SSCC sub-sampling method relies on the estimation of the population parameter $\tau$. However, this estimation will change depending on the application. This provides a natural way to calibrate the classifier for a different population of interest. The results show that by using the SSCC sub-sampling method, we greatly reduce the computational burden of
fitting the model on the high-dimensional data set. We are able to fit the models 180 times faster than if we used the full data set. When using only this sub-sample to fit the proposed model, predictive performance is comparable to using the full data set is achieved. This development allows for fitting next-generation local image regression models, which have recently shown promise in a variety of image analysis problems (Sweeney et al., 2013c,b; Shinohara et al., 2012), on hundreds or thousands of subjects observed at many visits.

In Chapter 4, we propose a functional linear model to estimate the current value of MTR using only the last six months of both the T1w and FLAIR voxel intensities. We provide two methods to estimate the regression parameter functions; the $PW$ method, a pointwise least squares approach; and the $GB$ method, a global penalized functional regression approach. The $PW$ method directly accounts for the covariance of the predictor functions by using a pointwise least squares approach to estimate the regression parameters. Additionally, in the case of sparse data, this method does not require the estimation of smooth trajectories as the $GB$ method does.

The $GB$ method does outperform the $PW$ method in both the simulation experiments but the methods provide comparable results in terms of cross-validated error for MTR trajectory estimation. Estimation of MTR using the $lag$FLM model with the recent six months of both the T1w and FLAIR voxel trajectories outperforms cross-sectional models that do not take temporal dependence of the voxel-level trajectories into account. An area of future work will be to explore the use of this $lag$FLM model to predict other quantitative imaging modalities such as T1w maps.
REFERENCES


Appendix A

Supplemental Material for Chapter 3

A.1 SuBLIME Map Estimation

Recently researchers have asserted that new or enlarging MS lesions are more likely to enhance (Gaitan et al., 2011). To account for this new finding, we include a covariate that is an indicator of new or enlarging lesion voxels to improve predictive ability in the proposed predictive model. Classification of voxels as belonging to new or enlarging MS lesions can be obtained manually by a radiologist. However, manual identification is extremely time consuming and prone to errors even for the most expert neuroradiologists in this field. Additionally, the process is subject to both intra-observer and inter-observer variability (Lladó et al., 2012a). Therefore, we propose the use of the SuBLIME method to estimate the probability that a voxel belongs to a new or enlarging lesion (Sweeney et al., 2013c).

The proposed modeling procedure is a two-step process. First, for each voxel used for fitting the model, we acquire a historical covariate that denotes whether the voxel belongs to a new or enlarging lesion. The second step is to use this covariate in the proposed logistic model. The covariate can either be estimated by obtaining a mask developed by a radiologist or from the SuBLIME method (Sweeney et al., 2013c) which estimates the probability that a voxel belongs to a new or enlarging lesion. We propose to employ the latter technique. The indicator of incidence as a binary variable is defined as, $W_i(v, t_{ij})$, which is equal to one if subject $i$ has new lesion incidence in voxel $v$ at time $t_{ij}$ and zero otherwise. The SuBLIME probability maps, $SP_i(v, t_{ij}) := Pr[W_i(v, t_{ij}) = 1]$, are the probabilities that voxel $v$ is with a new or enlarging lesion at time $t_{ij}$. As described in the previous section, these maps are obtained using longitudinal study information. The following is the model proposed by Sweeney et al. (2013c)
to estimate the probability maps of interest.

\[
\text{logit}[Pr\{W_i(v, t_{ij})\}] = \beta_0 + \beta_1 \Delta t_i + \beta_2 M_{i,1}(v, t_{ij}) + \beta_3 \Delta M_{i,1}(v, t_{ij}) \times \Delta t_i + \beta_5 M_{i,2}(v, t_{ij}) + \beta_6 \Delta M_{i,2}(v, t_{ij}) \times \Delta t_i + \beta_8 M_{i,3}(v, t_{ij}) + \beta_9 \Delta M_{i,3}(v, t_{ij}) \times \Delta t_i + \beta_{11} M_{i,4}(v, t_{ij}) + \beta_{12} \Delta M_{i,4}(v, t_{ij}) \times \Delta t_i
\]

The covariates in this model are defined as follows: \(M_{i,1} := T1w\) image for patient \(i\), \(M_{i,2} := T2w\) image for patient \(i\), \(M_{i,3} := FLAIR\) image for patient \(i\), \(M_{i,4} := PD\) image for patient \(i\), \(\Delta t_i := \) time in days between consecutive studies for patient \(i\), \(\Delta M_{i,k}(v, t_{ij}) = M_{i,k}(v, t_{ij}) - M_{i,k}(v, t_{ij-1})\) : subtraction image for patient \(i\) and imaging modality \(k\). By using subtraction images and the interaction with the time differences between scans, this model incorporates the longitudinal information.

A.2 Data Pre-Processing

The data are processed using the methods of Shinohara et al. (2012). The following is the description of the pre-processing steps presented in their paper:

Images were acquired on either a 1.5T (X scans) or a 3T (X scans) MR imaging scanner (Signa Excite HDxt; GE Healthcare, Milwaukee, Wisconsin) by using the body coil for transmission and an 8-channel receive coil array (Invivo, Gainesville, Florida) for signal-intensity detection. Depending on the platform, the sequence parameters differed. T1-weighted scans were obtained before contrast administration by using a 3D FSPGR sequence, with TR=9, TE=3.5ms, T1=450ms, FA=13°, and nominal VV=1-1.2 mm\(^3\). T2-weighted scans were acquired before contrast administration by using a 2D fast spin-echo sequence (TR = 5-6 seconds, TE \(\sim\) 120 ms, FA = 90°, VV = 1-2 mm\(^3\)).

An intravenous infusion of 0.1 mmol/kg of gadopentate dimeglumine (Magnevist; Bayer HealthCare, Leverkusen, Germany) via a power injector (Solaris; Medrad, Indiana, Pennsylvania) was administered to the subjects. T1-weighted scans were also obtained a median of 8 minutes (IQR, 6-15 minutes) after contrast injection by using either a 2D spin-echo sequence (TR=600 ms, TE = 16ms, FA=90°, VV=2 mm\(^3\)) or the same 3D FSPGR sequence used for pre-contrast T1-weighted scans. T2-weighted FLAIR scans were obtained at a median of 15 minutes (IQR, 14-45 minutes) after contrast injection. The T2-weighted FLAIR acquisition was conducted using either a 2D fast spin-echo sequence on the 1.5T scanner (TR=10 seconds, TE \(\sim\) 123 ms, TI=2250 ms, FA=90°, VV=3.5 mm\(^3\)) or a 3D variable FA sequence on the 3T scanner (TR=6 seconds, TE \(\sim\) 126 ms, TI \(\sim\) 1860 ms, VV=1 mm\(^3\)). Post contrast T2-weighted FLAIR scans facilitate detection of contrast-enhancing lesions without changing the signal in-
tensity of non-enhancing lesions so that the signal intensity within lesions on post-contrast scans is at least as great as that on precontrast scans Kataoka et al. (2009) Bagnato (2009).

Medical Image Processing Analysis and Visualization (http://mipav.cit.nih.gov) and Java Image Science Toolkit (http://nitrc.org/projects/jist) were used for processing the images. All statistical calculations and modeling were conducted by using the software environment R (version 3.1.0; R Foundation for Statistical Computing, Vienna, Austria).

All acquired volumes were rigidly registered to the pregadolinium T1-weighted volume and then rigidly aligned to the Montreal Neurologic Institute standard template. Nonparametric intensity-nonuniformity normalization was used to address scanner-related inhomogeneity Sled et al. (1998). All scans were interpolated to a voxel size of 1mm³. Skull and extracranial voxels were masked out by using a skull-stripping procedure Carass et al. (2007). The volume was eroded by 2mm in each direction to remove much of the residual extracerebral tissue. This should not remove voxels consisting white matter where lesions may occur. To more completely remove the normally enhancing meninges and to focus our attention on the white matter where most enhancing voxels are located, we removed all voxels below axial section 52 (the inferior temporal lobes) and above the axial section 156 (the top of the brain).
Appendix B

Supplemental Material for Chapter 4

B.1 Figures of the Simulated Data

Figure B.1: Observed data under simulation settings (I) - row one and (II) - row two.
Figure B.2: Smoothed predictor functions using 20 basis functions.

Figure B.3: Predicted response functions using the $PW$ estimation method (row one) and the $GB$ method (row two).

B.2 Additional Results for lagFHL Model with non-iid Measurement Error

B.2.1 Results for Dense Data with One Predictor (eSNR = 1)
Table B.1: Prediction Error for PW - Setting (I), One Predictor, eSNR = 1, Average Error Variance = 0.043

<table>
<thead>
<tr>
<th>n</th>
<th>(\Delta)</th>
<th>(\Delta_F)</th>
<th>(\text{rMSE}(\beta_0))</th>
<th>(\text{rMSE}(\beta_1))</th>
<th>(\text{rMSE}(Y)) (SE[\text{rMSE}(Y)])</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.4</td>
<td>2.6e-05</td>
<td>0.25</td>
<td>2.1e-03 (6.9e-04)</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.6</td>
<td>2.2e-05</td>
<td>0.43</td>
<td>2.4e-03 (9.4e-04)</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.4</td>
<td>1.1e-05</td>
<td>0.15</td>
<td>9.8e-04 (3.4e-04)</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.6</td>
<td>9.6e-06</td>
<td>0.26</td>
<td>1.1e-03 (4.1e-04)</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.4</td>
<td>5.5e-06</td>
<td>0.12</td>
<td>5.2e-04 (1.8e-04)</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.6</td>
<td>4.6e-06</td>
<td>0.19</td>
<td>5.6e-04 (2e-04)</td>
</tr>
</tbody>
</table>

Table B.2: Prediction Error for GB - Setting (I), One Predictor, eSNR=1 Average Error Variance = 0.043

<table>
<thead>
<tr>
<th>n</th>
<th>(\Delta)</th>
<th>(\Delta_F)</th>
<th>(\text{rMSE}(\beta_0))</th>
<th>(\text{rMSE}(\beta_1))</th>
<th>(\text{rMSE}(Y)) (SE[\text{rMSE}(Y)])</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.4</td>
<td>2.9e-05</td>
<td>0.068</td>
<td>9.9e-04 (4.3e-04)</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.6</td>
<td>2.4e-05</td>
<td>0.093</td>
<td>1.3e-03 (6.2e-04)</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.4</td>
<td>1.3e-05</td>
<td>0.029</td>
<td>4.3e-04 (1.7e-04)</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.6</td>
<td>1.1e-05</td>
<td>0.069</td>
<td>5.6e-04 (2.4e-04)</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.4</td>
<td>6.2e-06</td>
<td>0.016</td>
<td>2.3e-04 (7.4e-05)</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.6</td>
<td>5.1e-06</td>
<td>0.059</td>
<td>3e-04 (1.1e-04)</td>
</tr>
</tbody>
</table>

B.2.2 Results for Dense Data with One Predictor (eSNR = 5)

Table B.3: Prediction Error for PW - Setting (I), One Predictor, SNR = 5, Average Error Variance = 9e-03

<table>
<thead>
<tr>
<th>n</th>
<th>(\Delta)</th>
<th>(\Delta_F)</th>
<th>(\text{rMSE}(\beta_0))</th>
<th>(\text{rMSE}(\beta_1))</th>
<th>(\text{rMSE}(Y)) (SE[\text{rMSE}(Y)])</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.4</td>
<td>2.4e-05</td>
<td>0.13</td>
<td>1e-03 (5.2e-04)</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.6</td>
<td>1.9e-05</td>
<td>0.19</td>
<td>1e-03 (5.3e-04)</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.4</td>
<td>1e-05</td>
<td>0.1</td>
<td>4.9e-04 (2.7e-04)</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.6</td>
<td>8.6e-06</td>
<td>0.13</td>
<td>4.7e-04 (2.7e-04)</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.4</td>
<td>5.1e-06</td>
<td>9e-02</td>
<td>2.7e-04 (1.6e-04)</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.6</td>
<td>4.1e-06</td>
<td>0.11</td>
<td>2.6e-04 (1.5e-04)</td>
</tr>
</tbody>
</table>
Table B.4: Prediction Error for GB - Setting (I), One Predictor, eSNR = 5, Average Error Variance = 9e-03

<table>
<thead>
<tr>
<th>n</th>
<th>Δ</th>
<th>ΔF</th>
<th>rMSE(β₀)</th>
<th>rMSE(β₁)</th>
<th>rMSE(Y) (SE[rMSE(Y)])</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.4</td>
<td>2.7e-05</td>
<td>0.038</td>
<td>2.4e-04 (9.6e-05)</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.6</td>
<td>2.1e-05</td>
<td>0.084</td>
<td>3.1e-04 (1.4e-04)</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.4</td>
<td>1.2e-05</td>
<td>0.016</td>
<td>1e-04 (3.7e-05)</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.6</td>
<td>9.3e-06</td>
<td>0.072</td>
<td>1.3e-04 (5.1e-05)</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.4</td>
<td>5.7e-06</td>
<td>9.6e-03</td>
<td>5.6e-05 (1.6e-05)</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.6</td>
<td>4.4e-06</td>
<td>0.065</td>
<td>6.9e-05 (2.3e-05)</td>
</tr>
</tbody>
</table>

B.2.3 Results for Dense Data with Two Predictors (eSNR = 1)

Table B.5: Prediction Error for PW - Setting (I), Two Predictors, eSNR = 1, Average Error Variance = 0.058

<table>
<thead>
<tr>
<th>n</th>
<th>Δ</th>
<th>ΔF</th>
<th>rMSE(β₀)</th>
<th>rMSE(β₁)</th>
<th>rMSE(β₂)</th>
<th>rMSE(Y) (SE[rMSE(Y)])</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.4</td>
<td>3.2e-05</td>
<td>1</td>
<td>2.4</td>
<td>7.9e-03 (2e-02)</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.6</td>
<td>3e-05</td>
<td>3.7</td>
<td>3.3</td>
<td>0.011 (0.035)</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.4</td>
<td>1.3e-05</td>
<td>0.5</td>
<td>1.7</td>
<td>3.8e-03 (6.1e-03)</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.6</td>
<td>1.3e-05</td>
<td>0.68</td>
<td>1.5</td>
<td>4.4e-03 (4.4e-03)</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.4</td>
<td>6.4e-06</td>
<td>0.21</td>
<td>1.2</td>
<td>2.2e-03 (1.8e-03)</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.6</td>
<td>6.1e-06</td>
<td>0.63</td>
<td>1.4</td>
<td>2.9e-03 (5.6e-03)</td>
</tr>
</tbody>
</table>

76
Table B.6: Prediction Error for GB - Setting (I), Two Predictors, eSNR = 1 , Average Error Variance = 0.058

<table>
<thead>
<tr>
<th>n</th>
<th>Δ</th>
<th>Δ_F</th>
<th>rMSE(β_0)</th>
<th>rMSE(β_1)</th>
<th>rMSE(β_2)</th>
<th>rMSE(Y) (SE[rMSE(Y)])</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.4</td>
<td>3.5e-05</td>
<td>0.084</td>
<td>0.38</td>
<td>2e-03 ( 6.8e-04 )</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.6</td>
<td>3.4e-05</td>
<td>0.12</td>
<td>0.68</td>
<td>2.7e-03 ( 1e-03 )</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.4</td>
<td>1.5e-05</td>
<td>0.034</td>
<td>0.37</td>
<td>9.3e-04 ( 2.8e-04 )</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.6</td>
<td>1.4e-05</td>
<td>0.064</td>
<td>0.58</td>
<td>1.2e-03 ( 3.7e-04 )</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.4</td>
<td>7.3e-06</td>
<td>0.019</td>
<td>0.41</td>
<td>5.1e-04 ( 1.4e-04 )</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.6</td>
<td>6.9e-06</td>
<td>0.055</td>
<td>0.55</td>
<td>6.4e-04 ( 2e-04 )</td>
</tr>
</tbody>
</table>

B.2.4 Results for Sparse Data with One Predictor (eSNR = 1)

Table B.7: Prediction Error for PW - Setting (II), One Predictor, SNR = 1 , Average Error Variance = 0.043

<table>
<thead>
<tr>
<th>n</th>
<th>Δ</th>
<th>Δ_F</th>
<th>rMSE(β_0)</th>
<th>rMSE(β_1)</th>
<th>rMSE(Y) (SE[rMSE(Y)])</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.4</td>
<td>4.2e-05</td>
<td>1.4</td>
<td>0.015 ( 7.1e-03 )</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.6</td>
<td>4e-05</td>
<td>2.5</td>
<td>0.015 ( 7.6e-03 )</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.4</td>
<td>1.6e-05</td>
<td>0.87</td>
<td>8.9e-03 ( 4.9e-03 )</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.6</td>
<td>1.6e-05</td>
<td>1.6</td>
<td>9.8e-03 ( 2e-02 )</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.4</td>
<td>8e-06</td>
<td>0.55</td>
<td>6.1e-03 ( 5.2e-03 )</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.6</td>
<td>7.9e-06</td>
<td>1.4</td>
<td>6e-03 ( 5.1e-03 )</td>
</tr>
</tbody>
</table>
Table B.8: Prediction Error for GB - Setting (II), One Predictor, eSNR = 1, Average Error Variance = 0.043

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.4</td>
<td>4.1e-05</td>
<td>0.27</td>
<td>7.8e-03 (5.6e-03)</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.6</td>
<td>3.9e-05</td>
<td>0.38</td>
<td>8.5e-03 (4.9e-03)</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.4</td>
<td>1.7e-05</td>
<td>0.13</td>
<td>5.0e-03 (6.2e-03)</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.6</td>
<td>1.5e-05</td>
<td>0.19</td>
<td>5.1e-03 (5.5e-03)</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.4</td>
<td>8.4e-06</td>
<td>0.081</td>
<td>3.8e-03 (4.1e-03)</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.6</td>
<td>7.7e-06</td>
<td>0.13</td>
<td>3.6e-03 (3.8e-03)</td>
</tr>
</tbody>
</table>

B.2.5 Results for Sparse Data with One Predictor (eSNR = 5)

Table B.9: Prediction Error for PW - Setting (II), One Predictor, eSNR = 5, Average Error Variance = 9e-03

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.4</td>
<td>3.4e-05</td>
<td>1.3</td>
<td>1e-02 (6.1e-03)</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.6</td>
<td>3.1e-05</td>
<td>2.3</td>
<td>9.8e-03 (5.6e-03)</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.4</td>
<td>1.3e-05</td>
<td>0.74</td>
<td>6.7e-03 (5.4e-03)</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.6</td>
<td>1.2e-05</td>
<td>1.3</td>
<td>6.4e-03 (6.6e-03)</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.4</td>
<td>6.6e-06</td>
<td>0.52</td>
<td>5e-03 (5.6e-03)</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.6</td>
<td>6.1e-06</td>
<td>0.99</td>
<td>4.5e-03 (4.6e-03)</td>
</tr>
</tbody>
</table>

78
Table B.10: Prediction Error for GB - Setting (II), One Predictor, eSNR = 5, Average Error Variance = 9e-03

<table>
<thead>
<tr>
<th>n</th>
<th>Δ</th>
<th>Δ_F</th>
<th>rMSE(β₀)</th>
<th>rMSE(β₁)</th>
<th>rMSE(Y) (SE[rMSE(Y)])</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.4</td>
<td>3.1e-05</td>
<td>0.22</td>
<td>5.1e-03 (6.6e-03)</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.6</td>
<td>2.6e-05</td>
<td>0.3</td>
<td>4.5e-03 (5.9e-03)</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.4</td>
<td>1.3e-05</td>
<td>0.16</td>
<td>4e-03 (7.8e-03)</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.6</td>
<td>1e-05</td>
<td>0.21</td>
<td>3.4e-03 (4.9e-03)</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.4</td>
<td>6.1e-06</td>
<td>0.12</td>
<td>3.2e-03 (4.3e-03)</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.6</td>
<td>5.1e-06</td>
<td>0.17</td>
<td>2.8e-03 (4.1e-03)</td>
</tr>
</tbody>
</table>

B.2.6 Results for Sparse Data with Two Predictors (eSNR = 1)

Table B.11: Prediction Error for PW - Setting (II), Two Predictors, eSNR = 1, Average Error Variance = 0.058

<table>
<thead>
<tr>
<th>n</th>
<th>Δ</th>
<th>Δ_F</th>
<th>rMSE(β₀)</th>
<th>rMSE(β₁)</th>
<th>rMSE(β₂)</th>
<th>rMSE(Y) (SE[rMSE(Y)])</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.4</td>
<td>4.7e-05</td>
<td>1.1</td>
<td>2.8</td>
<td>0.025 (0.063)</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.6</td>
<td>4.8e-05</td>
<td>3.6</td>
<td>4.6</td>
<td>0.036 (0.15)</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.4</td>
<td>1.9e-05</td>
<td>0.49</td>
<td>1.7</td>
<td>0.013 (4.4e-03)</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.6</td>
<td>1.9e-05</td>
<td>0.64</td>
<td>1.6</td>
<td>0.015 (5.6e-03)</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.4</td>
<td>9.4e-06</td>
<td>0.35</td>
<td>1.5</td>
<td>8.6e-03 (4.3e-03)</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.6</td>
<td>9.7e-06</td>
<td>0.58</td>
<td>1.6</td>
<td>9.8e-03 (5e-03)</td>
</tr>
</tbody>
</table>
Table B.12: Prediction Error for GB - Setting (II), Two Predictors, eSNR = 1, Average Error Variance = 0.058

<table>
<thead>
<tr>
<th>n</th>
<th>Δ</th>
<th>ΔF</th>
<th>rMSE($\beta_0$)</th>
<th>rMSE($\beta_1$)</th>
<th>rMSE($\beta_2$)</th>
<th>rMSE(Y) (SE[rMSE(Y)])</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.4</td>
<td>4.6e-05</td>
<td>0.33</td>
<td>0.63</td>
<td>0.012 ( 0.017 )</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.6</td>
<td>5.2e-05</td>
<td>0.49</td>
<td>1.2</td>
<td>0.017 ( 0.056 )</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.4</td>
<td>1.8e-05</td>
<td>0.13</td>
<td>0.35</td>
<td>6.6e-03 ( 3.3e-03 )</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.6</td>
<td>1.9e-05</td>
<td>0.19</td>
<td>0.69</td>
<td>7.5e-03 ( 2.9e-03 )</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.4</td>
<td>9.9e-06</td>
<td>8e-02</td>
<td>0.31</td>
<td>4.8e-03 ( 3.4e-03 )</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.6</td>
<td>1e-05</td>
<td>0.13</td>
<td>0.6</td>
<td>5e-03 ( 2.6e-03 )</td>
</tr>
</tbody>
</table>

B.3 Additional Results for lagFHL Model with iid Measurement Error

B.3.1 Results for Dense Data with One Predictor (eSNR = 1)

Table B.13: Prediction Error for PW - Setting (I), One Predictor, iid Error, eSNR = 1, Average Error Variance = 0.043

<table>
<thead>
<tr>
<th>n</th>
<th>Δ</th>
<th>ΔF</th>
<th>rMSE($\beta_0$)</th>
<th>rMSE($\beta_1$)</th>
<th>rMSE(Y) (SE[rMSE(Y)])</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.4</td>
<td>5.1e-05</td>
<td>0.36</td>
<td>8.5e-03 ( 2.4e-03 )</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.6</td>
<td>4.8e-05</td>
<td>0.68</td>
<td>9.2e-03 ( 2.8e-03 )</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.4</td>
<td>2.1e-05</td>
<td>0.19</td>
<td>3.7e-03 ( 1.1e-03 )</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.6</td>
<td>2e-05</td>
<td>0.35</td>
<td>4e-03 ( 1.2e-03 )</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.4</td>
<td>1.1e-05</td>
<td>0.14</td>
<td>1.9e-03 ( 5.4e-04 )</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.6</td>
<td>9.9e-06</td>
<td>0.22</td>
<td>2e-03 ( 6.2e-04 )</td>
</tr>
</tbody>
</table>
Table B.14: Prediction Error for GB - Setting (I), One Predictor, idd Error, eSNR = 1, Average Error Variance = 0.043

<table>
<thead>
<tr>
<th>n</th>
<th>Δ</th>
<th>Δ_F</th>
<th>(rMSE(\beta_0))</th>
<th>(rMSE(\beta_1))</th>
<th>(rMSE(Y)) (SE[(rMSE(Y))])</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.4</td>
<td>2.7e-05</td>
<td>0.046</td>
<td>5.1e-04 ( 2.2e-04 )</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.6</td>
<td>2.3e-05</td>
<td>0.061</td>
<td>6.8e-04 ( 3.1e-04 )</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.4</td>
<td>1.2e-05</td>
<td>2e-02</td>
<td>2.2e-04 ( 7.7e-05 )</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.6</td>
<td>9.9e-06</td>
<td>0.039</td>
<td>2.9e-04 ( 1.2e-04 )</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.4</td>
<td>6e-06</td>
<td>0.011</td>
<td>1.3e-04 ( 4.2e-05 )</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.6</td>
<td>5.1e-06</td>
<td>0.033</td>
<td>1.6e-04 ( 6.1e-05 )</td>
</tr>
</tbody>
</table>

B.3.2 Results for Dense Data with One Predictor (eSNR = 5)

Table B.15: Prediction Error for PW - Setting (I), One Predictor, idd Error, eSNR = 5, Average Error Variance = 9e-03

<table>
<thead>
<tr>
<th>n</th>
<th>Δ</th>
<th>Δ_F</th>
<th>(rMSE(\beta_0))</th>
<th>(rMSE(\beta_1))</th>
<th>(rMSE(Y)) (SE[(rMSE(Y))])</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.4</td>
<td>3.1e-05</td>
<td>0.14</td>
<td>3.7e-03 ( 2e-03 )</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.6</td>
<td>2.7e-05</td>
<td>0.2</td>
<td>3.6e-03 ( 1.9e-03 )</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.4</td>
<td>1.3e-05</td>
<td>0.099</td>
<td>1.7e-03 ( 1e-03 )</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.6</td>
<td>1.1e-05</td>
<td>0.13</td>
<td>1.6e-03 ( 9.8e-04 )</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.4</td>
<td>6.6e-06</td>
<td>0.086</td>
<td>9.1e-04 ( 5.2e-04 )</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.6</td>
<td>5.7e-06</td>
<td>0.098</td>
<td>8.4e-04 ( 5.4e-04 )</td>
</tr>
</tbody>
</table>
Table B.16: Prediction Error for GB -Setting (I), One Predictor, idd Error, eSNR = 5, Average Error Variance = 9e-03

<table>
<thead>
<tr>
<th>n</th>
<th>Δ</th>
<th>Δ F</th>
<th>rMSE(β₀)</th>
<th>rMSE(β₁)</th>
<th>rMSE(Y) (SE[rMSE(Y)])</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.4</td>
<td>2.6e-05</td>
<td>0.031</td>
<td>1.3e-04 ( 4.9e-05 )</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.6</td>
<td>2.2e-05</td>
<td>0.054</td>
<td>1.6e-04 ( 6.8e-05 )</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.4</td>
<td>1.1e-05</td>
<td>0.014</td>
<td>5.6e-05 ( 1.8e-05 )</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.6</td>
<td>9.3e-06</td>
<td>0.043</td>
<td>7.1e-05 ( 2.6e-05 )</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.4</td>
<td>5.7e-06</td>
<td>7.5e-03</td>
<td>3.2e-05 ( 9.4e-06 )</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.6</td>
<td>4.7e-06</td>
<td>0.036</td>
<td>3.8e-05 ( 1.3e-05 )</td>
</tr>
</tbody>
</table>

B.3.3 Results for Dense Data with Two Predictors (eSNR = 1)

Table B.17: Prediction Error for PW -Setting (I), Two Predictors, idd Error, eSNR = 1, Average Error Variance = 0.058

<table>
<thead>
<tr>
<th>n</th>
<th>Δ</th>
<th>Δ F</th>
<th>rMSE(β₀)</th>
<th>rMSE(β₁)</th>
<th>rMSE(β₂)</th>
<th>rMSE(Y) (SE[rMSE(Y)])</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.4</td>
<td>6.5e-05</td>
<td>0.44</td>
<td>1.3</td>
<td>0.013 ( 3e-03 )</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.6</td>
<td>6.5e-05</td>
<td>0.83</td>
<td>1.4</td>
<td>0.016 ( 5.2e-03 )</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.4</td>
<td>2.6e-05</td>
<td>0.27</td>
<td>1.1</td>
<td>6.2e-03 ( 1.2e-03 )</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.6</td>
<td>2.6e-05</td>
<td>0.42</td>
<td>1.1</td>
<td>7.5e-03 ( 1.6e-03 )</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.4</td>
<td>1.3e-05</td>
<td>0.18</td>
<td>0.97</td>
<td>3.6e-03 ( 6.6e-04 )</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.6</td>
<td>1.2e-05</td>
<td>0.33</td>
<td>1</td>
<td>4.3e-03 ( 9e-04 )</td>
</tr>
</tbody>
</table>
### B.3.4 Results for Sparse Data with One Predictor (eSNR = 1)

Table B.19: Prediction Error for \( PW \) -Setting (II), One Predictor, idd Error, eSNR = 1, Average Error Variance = 0.043

<table>
<thead>
<tr>
<th>( n )</th>
<th>( \Delta )</th>
<th>( \Delta_F )</th>
<th>( rMSE(\beta_0) )</th>
<th>( rMSE(\beta_1) )</th>
<th>( rMSE(\beta_2) )</th>
<th>( rMSE(Y) ) ((SE[rMSE(Y)]))</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.4</td>
<td>3.9e-05</td>
<td>1.5</td>
<td>0.014</td>
<td>0.014 (6.3e-03)</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.6</td>
<td>3.8e-05</td>
<td>2.8</td>
<td>0.015</td>
<td>0.015 (8.6e-03)</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.4</td>
<td>1.8e-05</td>
<td>0.9</td>
<td>8.8e-03</td>
<td>8.8e-03 (7.2e-03)</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.6</td>
<td>1.7e-05</td>
<td>1.5</td>
<td>8.8e-03</td>
<td>8.8e-03 (7.2e-03)</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.4</td>
<td>8.1e-06</td>
<td>0.55</td>
<td>8.5e-03</td>
<td>8.5e-03 (0.043)</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.6</td>
<td>7.9e-06</td>
<td>1.3</td>
<td>8.8e-03</td>
<td>8.8e-03 (0.055)</td>
</tr>
</tbody>
</table>
Table B.20: Prediction Error for \textit{GB} -Setting (II), One Predictor, idd Error, eSNR = 1, Average Error Variance = 0.043

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\Delta$</th>
<th>$\Delta_F$</th>
<th>$rMSE(\beta_0)$</th>
<th>$rMSE(\beta_1)$</th>
<th>$rMSE(Y)$ (SE[rMSE(Y)])</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.4</td>
<td>3.7e-05</td>
<td>0.24</td>
<td>7.2e-03 (5.3e-03)</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.6</td>
<td>3.6e-05</td>
<td>0.31</td>
<td>8.1e-03 (5.6e-03)</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.4</td>
<td>1.7e-05</td>
<td>0.12</td>
<td>4.9e-03 (5.9e-03)</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.6</td>
<td>1.6e-05</td>
<td>0.16</td>
<td>4.8e-03 (5.4e-03)</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.4</td>
<td>8e-06</td>
<td>0.078</td>
<td>5.7e-03 (4e-02)</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.6</td>
<td>7.6e-06</td>
<td>0.13</td>
<td>6e-03 (0.053)</td>
</tr>
</tbody>
</table>

B.3.5 Results for Sparse Data with One Predictor (eSNR = 5)

Table B.21: Prediction Error for \textit{PW} -Setting (II), One Predictor, idd Error, eSNR = 5, Average Error Variance = 9e-03

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\Delta$</th>
<th>$\Delta_F$</th>
<th>$rMSE(\beta_0)$</th>
<th>$rMSE(\beta_1)$</th>
<th>$rMSE(Y)$ (SE[rMSE(Y)])</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.4</td>
<td>3.3e-05</td>
<td>1.2</td>
<td>1e-02 (5.4e-03)</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.6</td>
<td>3e-05</td>
<td>1.9</td>
<td>1e-02 (7.5e-03)</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.4</td>
<td>1.4e-05</td>
<td>0.74</td>
<td>7e-03 (8.6e-03)</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.6</td>
<td>1.3e-05</td>
<td>0.95</td>
<td>6.3e-03 (6.1e-03)</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.4</td>
<td>7e-06</td>
<td>0.54</td>
<td>7.3e-03 (0.041)</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.6</td>
<td>6.5e-06</td>
<td>0.87</td>
<td>7.1e-03 (0.052)</td>
</tr>
</tbody>
</table>
Table B.22: Prediction Error for GB -Setting(II), One Predictor, idd Error, eSNR = 5, Average Error Variance = 9e-03

<table>
<thead>
<tr>
<th>n</th>
<th>∆</th>
<th>∆_F</th>
<th>rMSE(β_0)</th>
<th>rMSE(β_1)</th>
<th>rMSE(Y) (SE[rMSE(Y)])</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.4</td>
<td>3e-05</td>
<td>0.21</td>
<td>4.6e-03 ( 4.9e-03 )</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.6</td>
<td>2.5e-05</td>
<td>0.25</td>
<td>4.3e-03 ( 4.2e-03 )</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.4</td>
<td>1.3e-05</td>
<td>0.14</td>
<td>3.9e-03 ( 6.1e-03 )</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.6</td>
<td>1.1e-05</td>
<td>0.19</td>
<td>3.4e-03 ( 6e-03 )</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.4</td>
<td>6.3e-06</td>
<td>0.12</td>
<td>5.2e-03 ( 0.038 )</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.6</td>
<td>5.5e-06</td>
<td>0.17</td>
<td>5.4e-03 ( 0.052 )</td>
</tr>
</tbody>
</table>

B.3.6 Results for Sparse Data with Two Predictors (eSNR = 1)

Table B.23: Prediction Error for PW -Setting(II), Two Predictors, idd Error, eSNR = 1, Average Error Variance = 0.058

<table>
<thead>
<tr>
<th>n</th>
<th>∆</th>
<th>∆_F</th>
<th>rMSE(β_0)</th>
<th>rMSE(β_1)</th>
<th>rMSE(β_2)</th>
<th>rMSE(Y) (SE[rMSE(Y)])</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.4</td>
<td>4.5e-05</td>
<td>0.88</td>
<td>2.1</td>
<td>0.021 ( 0.011 )</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.6</td>
<td>4.6e-05</td>
<td>270</td>
<td>110</td>
<td>0.63 ( 13 )</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.4</td>
<td>1.9e-05</td>
<td>0.46</td>
<td>1.6</td>
<td>0.012 ( 4.1e-03 )</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.6</td>
<td>2e-05</td>
<td>1.2</td>
<td>1.9</td>
<td>0.015 ( 0.023 )</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.4</td>
<td>9.7e-06</td>
<td>0.33</td>
<td>1.4</td>
<td>9.4e-03 ( 0.016 )</td>
</tr>
<tr>
<td>100</td>
<td>0.4</td>
<td>0.6</td>
<td>1e-05</td>
<td>6.5</td>
<td>13</td>
<td>0.025 ( 0.32 )</td>
</tr>
</tbody>
</table>
Recall that \( \beta_p(s, t) \) is represented as \( \beta_p(s, t) = \sum_{k=1}^{K} B_{sp,k}(s) \alpha_{pk}(t) \), for fixed \( K \), fixed basis functions \( B_{sp,k}(s) \) and unknown \( \alpha_{pk}(t) \). We estimate \( \hat{\beta}_p(s, t) = \sum_{k=1}^{K} B_{sp,k}(s) \hat{\alpha}_{pk}(t) \) with \( \hat{\alpha}_{pk}(t) \) as defined in equation (4.7). We will show that this is a consistent estimator (i.e., as \( \Delta \to 0 \) as \( n \to \infty \)). As a first step, we show that \( \hat{\alpha}_{pk}(t) \) is a consistent estimator of \( \alpha_{pk}(s, t) \). For this, the following assumptions are needed:

[A1] the estimators for the functional mean and covariates are consistent, i.e.

\[
\begin{align*}
    \sup_{t \in T} |\hat{\mu}_p(t) - \mu_p(t)| &= O_p(n^{-\alpha}) \\
    \sup_{t \in T} |\hat{\mu}_Y(t) - \mu_Y(t)| &= O_p(n^{-\alpha}) \\
    \sup_{s_1, s_2 \in T} |\hat{G}_{X_pX_q}(s_1, s_2) - G_{X_pX_q}(s_1, s_2)| &= O_p(n^{-\alpha}) \\
    \sup_{s_1, s_2 \in T} |\hat{G}_{X_pY}(s_1, s_2) - G_{X_pY}(s_1, s_2)| &= O_p(n^{-\alpha})
\end{align*}
\]

where \( 0 < \alpha < 1 \),

[A2] \( B_{s1,k}(s) \) and \( B_{s2,k}(s) \) are uniformly bounded, i.e. \( \sup_{s \in [0, \Delta]} |B_{s1,k}(s)| ds \leq C < \infty \) and similarly for \( B_{s2,k}(s) \),

[A3] \( \lambda_1 \) and \( \lambda_2 \) are bounded, and

[A4] the matrix

\[
G(t) = \begin{bmatrix}
    G_{11}(t) & G_{12}(t) \\
    G_{21}(t) & G_{22}(t)
\end{bmatrix}
\]

is invertible for all \( t \), where \( G_{pq}(t) = \{G_{pq,kt}(t)\}_{kl} \) is a \( K \times K \) matrix,

\[
G_{pq,kt}(t) = \int_0^\Delta \int_0^\Delta G_{X_pX_q}(t - s_1, t - s_2) B_{sp,t}(s_1) B_{sq,k}(s_2) ds_1 ds_2,
\]

and \( G_{X_pX_q} \) is the covariance between \( X_p \) and \( X_q \) for \( p, q \in \{1, 2\} \).
Lemma 1 Assume \([A1] - [A4]\), then for each \(k \in \{1, \ldots, K\}\), \(\sup_{t \in T} |\hat{\alpha}_{pk}(t) - \alpha_{pk}(t)| \xrightarrow{p} 0\) as \(n \to \infty\).

Proof. By \([A4]\) and the definition of Model 4.2, \(\alpha(t) = \{G(t)\}^{-1} G_Y(t)\). We will show that \(\hat{\alpha}(t) = \{\hat{G}(t)\}^{-1} \hat{G}_Y(t)\) is a consistent estimator. First, note that consistency of \(\hat{G}_{X_pX_p}(t)\) and \(\hat{G}_{pY}(t)\) follows from the consistency of \(\hat{G}_{X_pX_p}(t)\) and \(\hat{G}_{X_pY}(t)\) (assumption A1). In order to prove this, it is observed that

\[
\left| \hat{G}_{pY}(t) - G_{pY}(t) \right| = \left| \int_0^\Delta B_{sp,k}(s) \left( \hat{G}_{X_pY}(t-s,t) - G_{X_pY}(t-s,t) \right) ds \right| \\
\leq \Delta \cdot \sup_{s \in [0,\Delta]} \left| B_{sp,k}(s) \right| \cdot \sup_{s \in [0,\Delta]} \left| \hat{G}_{X_pY}(t-s,t) - G_{X_pY}(t-s,t) \right| \\
\leq \Delta \cdot C \cdot \sup_{s \in [0,\Delta]} \left| \hat{G}_{X_pY}(t-s,t) - G_{X_pY}(t-s,t) \right| .
\]

(B.3)

This result leads to

\[
\sup_{t \in [\Delta, T]} \left| \hat{G}_{pY}(t) - G_{pY}(t) \right| \leq \Delta \cdot C \cdot \sup_{s_1, s_2 \in T} \left| \hat{G}_{X_pY}(s_1, s_2) - G_{X_pY}(s_1, s_2) \right| ,
\]

which implies that \(\sup_{t \in T} |\hat{G}_{pY}(t) - G_{pY}(t)| = O_p(n^{-\alpha})\) by assumption \([A1]\). A similar result can be shown for \(\hat{G}_{X_pX_p}(t)\). Hence, the corresponding vectors \(\hat{G}_{11}(t)\) and \(\hat{G}_{12}(t)\) and matrices \(\hat{G}_{11}(t)\), \(\hat{G}_{12}(t)\), and \(\hat{G}_{22}(t)\) are consistent since each element is consistent. Furthermore, if we let

\[
\hat{G}(t) = \begin{bmatrix} \hat{G}_{11}(t) & \hat{G}_{12}(t) \\ \hat{G}_{21}(t) & \hat{G}_{22}(t) \end{bmatrix} + \begin{bmatrix} \frac{\Delta_n}{n} I_K & 0 \\ 0 & \frac{\Delta_n}{n} I_K \end{bmatrix}
\]

(B.5)

then \(\hat{G}(t)\) is a consistent estimator of \(\{G(t)\}\) and by condition \([A4]\), \(\hat{G}(t)\) is a consistent estimator of \(\{G(t)\}^{-1}\). Therefore, \(\hat{\alpha}(t) = \{\hat{G}(t)\}^{-1} \hat{G}_Y(t)\) is a consistent estimator of \(\alpha(t) = \{G(t)\}^{-1} G_Y(t)\). \(\square\)

Proposition 1. Assume \([A1] - [A4]\), then \(\hat{\beta}_p(s,t)\) is consistent, i.e. \(\sup_{s \in [0,\Delta], t \in T} |\hat{\beta}_p(s,t) - \beta_p(s,t)| \xrightarrow{p} 0\) as \(n \to \infty\).

Proof. In order to prove consistency of \(\hat{\beta}_p(s,t)\), we must show that

\[
\sup_{s \in [0,\Delta], t \in T} |\hat{\beta}_p(s,t) - \beta_p(s,t)| \xrightarrow{p} 0.
\]

(B.6)

By definition of the estimator \(\left| \hat{\beta}_p(s,t) - \beta_p(s,t) \right| \leq \sum_{k=1}^K |B_{sp,k}(s)| \cdot |\hat{\alpha}_{pk}(t) - \alpha_{pk}(t)|\). Hence,

\[
\sup_{s \in [0,\Delta], t \in T} |\hat{\beta}_p(s,t) - \beta_p(s,t)| \leq \sum_{k=1}^K C \cdot \sup_{t \in T} |\hat{\alpha}_{pk}(t) - \alpha_{pk}(t)| .
\]

(B.7)

As it is shown in lemma 1, we have that \(\sup_{t \in T} |\hat{\alpha}_{pk}(t) - \alpha_{pk}(t)| \xrightarrow{p} 0\) as \(n \to \infty\). Hence, \(\sup_{s \in [0,\Delta], t \in T} |\hat{\beta}_p(s,t) - \beta_p(s,t)| \xrightarrow{p} 0\), which proves the result. \(\square\)
One Class of Functions that Satisfy Condition [A4]

As an illustration of one class of functions that satisfy condition [A4] for a given set of basis \( \{B_{sp,k}\}_k \), let \( G_{X_1,X_1}(s_1,s_2) \) have an eigen-expansion \( \sum_{l=1}^{L_1} \eta_{1l} \cdot \psi_{1l}(s_1)\psi_{1l}(s_2) \) (with \( \eta_{1l} > 0 \) and \( \{\psi_{1l}\} \) the set of orthonormal eigenbasis) for a fixed truncation parameter \( L_1 \) and similarly for \( G_{X_2,X_2}(s_1,s_2) \). Condition [A4] is satisfied if \( X_1(t) \) and \( X_2(t) \) are independent processes, and for all \( t \) there does not exist a non-trivial linear combination of \( \{B_{s1,k}(s)\}_k \) which lie in the nullspace of \( \{\psi_{1l}(t-s)\}_l \) where \( s \in [0,\Delta] \), and similarly for \( \{B_{s2,k}(s)\}_k \). To show this, we use the following results.

**Result 1** If for each \( t \) there does not exist a non-trivial linear combination of \( \{B_{s1,k}(s)\}_k \) which lie in the nullspace of \( \{\psi_{1l}(t-s)\}_l \) where \( s \in [0,\Delta] \), then \( C_1(t) = \left[ \int_0^\Delta \psi_{1l}(t-s)B_{s1,k}(s)\,ds \right]_{l,k} \) and \( C_2(t) = \left[ \int_0^\Delta \psi_{2l}(t-s)B_{s2,k}(s)\,ds \right]_{l,k} \) have rank equal to \( K \) for all \( t \).

**Proof.**

We prove Result 1 by contradiction. Assume that the rank of \( C_1(t) \) is not \( K \), so there exists a non-trivial vector \( a \) such that \( C_1(t) a = 0 \) for some \( t \). Then, \( \sum_k^K a_k \int_0^\Delta \psi_{1l}(t-s)B_{s1,k}(s)\,ds = 0 \) for all \( l \), which implies that \( \int_0^\Delta \psi_{1l}(t-s) \left( \sum_k^K a_k B_{s1,k}(s) \right)\,ds = 0 \). This last equality indicates that \( \sum_k^K a_k B_{s1,k}(s) \) is orthogonal to the functions \( \{\psi_{1l}(t-s)\}_l \). Therefore, \( \sum_k^K a_k B_{s1,k}(s) \) is in the nullspace of \( \{\psi_{1l}(t-s)\}_l \), which contradicts the initial claim. Hence, \( C_1(t) \) (of dimension \( L_1 \times K \)) is rank \( K \) and similarly for \( C_2(t) \). \( \square \)

**Result 2** \( G_{11}(t) = C_1(t)^\top D_1 C_1(t) \) where \( D_1 = \text{diag}(\eta_{11},\eta_{12},\cdots,\eta_{1L_1}) \) as defined above, with a similar expression for \( G_{22}(t) \).

**Proof.**

Observe that the \((k,k')\)th entry of \( G_{11}(t) \) is given by,

\[
G_{11,k,k'}(t) = \int_0^\Delta \int_0^\Delta \left[ \int \eta_{1l} \psi_{1l}(t-s_1)\psi_{1l}(t-s_2) \right] B_{s1,k}(s_1)B_{s1,k'}(s_2)\,ds_1\,ds_2
\]

\[
= \int \eta_{1l} \left[ \int \psi_{1l}(t-s_1)B_{s1,k}(s_1)\,ds_1 \right] \cdot \left[ \int \psi_{1l}(t-s_2)B_{s1,k'}(s_2)\,ds_2 \right]
\]

\[
= \int \eta_{1l} C_{1,k}(t) C_{1,k'}(t) = C_{1,k}(t)^\top D_1 C_{1,k'}(t),
\]

where \( C_{1,k}(t) \) is the \( k \)-th column of \( C_1(t) \). Therefore, \( G_{11}(t) = C_1(t)^\top D_1 C_1(t) \). \( \square \)

Using these results we show that condition [A4] is satisfied. Note that since \( X_1(t) \) and \( X_2(t) \) are independent processes, the covariances \( G_{12}(t) = G_{21}(t) = 0 \), which makes \( G(t) \) a block diagonal matrix with \( G_{11}(t) \) and \( G_{22}(t) \) in the diagonal. To show \( G(t) \) is non-singular it is sufficient to show that \( G_{11}^{-1}(t) \) and \( G_{22}^{-1}(t) \) exist.

To do this, note that by Result 1, \( C_1(t) \) (of dimension \( L_1 \times K \)) has rank \( K \). Therefore, for the \( L_1 \times L_1 \) matrix \( D_1 \), Silvester’s rank inequality yields, \( \text{rank} \left[ D_1^{1/2} C_1 \right] \geq \text{rank} \left[ D_1^{1/2} \right] + \text{rank} \left[ C_1 \right] - 88 \).
\[ L_1 = K. \text{ Therefore, the rank of } G_{11}(t) \text{ is } \text{rank} \left[ C_1^T D_1 C_1 \right] = \text{rank} \left[ (D_1^{1/2} C_1)^T (D_1^{1/2} C_1) \right] = \text{rank} \left[ D_1^{1/2} C_1 \right] = K. \text{ Hence, the } K \times K \text{ matrix } G_{11}(t) \text{ is non-singular. The same result holds for } G_{22}(t). \text{ Thus, the inverse of } G(t) \text{ is a block diagonal matrix with } G_{11}^{-1}(t) \text{ and } G_{22}^{-1}(t) \text{ in the diagonal, and so } G(t) \text{ is non-singular.} \]

**B.4.2 Consistency of Response Estimators**

**Proposition 2.** Let \( X_1^*(t), X_2^*(t) \) and \( Y^*(t) \) be a set of predictors and corresponding response function with estimator \( \hat{Y}^*(t) \) given by equation (4.10). If assumptions [A1]-[A4] hold and \( X_1^*(\cdot) \) and \( X_2^*(\cdot) \) are absolutely integrable, then the estimator \( \hat{Y}^*(t) \) is consistent as \( n \to \infty \), i.e.

\[
\sup_{t \in T} \left| \hat{Y}^*(t) - E[Y^*(t)|X_1^*(\cdot), X_2^*(\cdot)] \right| \overset{P}{\to} 0,
\]

where \( E[Y^*(t)|X_1^*(\cdot), X_2^*(\cdot)] = \beta_0(t) + \int_0^\Delta \beta_1(s,t)X_1^*(t-s)ds + \int_0^\Delta \beta_2(s,t)X_2^*(t-s)ds. \)

**Proof.** Define \( f^*(t) = E[Y^*(t)|X_1^*(\cdot), X_2^*(\cdot)] \) and note that

\[
\left| \hat{Y}^*(t) - f^*(t) \right| \leq \left| \hat{\beta}_0(t) - \beta_0(t) \right| + \sup_{s \in [0,\Delta]} \left| \hat{\beta}_1(s,t) - \beta_1(s,t) \right| \cdot \int_{t-\Delta}^t |X_1^*(s)|ds
+ \sup_{s \in [0,\Delta], t \in T} \left| \hat{\beta}_2(s,t) - \beta_2(s,t) \right| \cdot \int_0^t |X_2^*(s)|ds.
\]

(B.9)

Hence, due to integrability of \( X_1^* \) and \( X_2^* \), consistency of \( \hat{Y}^*(t) \) follows from consistency of \( \hat{\beta}_0, \hat{\beta}_1 \) and \( \hat{\beta}_2 \). Recall that \( \beta_0(t) = \mu_Y(t) \) and hence the estimator is consistent by condition 2 in assumption [A1]. Proposition 1 ensures that \( \hat{\beta}_1(s,t) \) and \( \hat{\beta}_2(s,t) \) are consistent. Hence, \( \sup_{t \in T} \left| \hat{Y}^*(t) - f^*(t) \right| \overset{P}{\to} 0 \), which concludes the proof. \( \Box \)

**Proposition 3.** Let \( X_1^*(t), X_2^*(t) \) and \( Y^*(t) \) be a set of predictors and corresponding response function with estimator \( \hat{Y}^*(t) \) given by equation (4.11). Assume [A1]-[A4] hold and

[B1] the estimators for the functional eigenvectors and scores for the KL decomposition of \( X_1^*(t) \) and \( X_2^*(t) \) satisfy

\[
\sup_{t \in T} \left| \hat{\psi}_{1l}(t) - \psi_{1l}(t) \right| = O_p(n^{-\alpha})
\]

\[
\sup_{t \in T} \left| \hat{\zeta}_{1l}^* - \zeta_{1l}^* \right| \overset{P}{\to} 0 \text{ as } n \to \infty \quad \text{ (B.11)}
\]

for \( l \geq 1 \) and with similar conditions for \( \hat{\psi}_{2l} \) and \( \hat{\zeta}_{2l}^* \).

[B2] Assume that \( \omega_{1l} = 0 \) for all \( l > L_1 \) for some \( L_1 \). Similarly assume \( \omega_{2l} = 0 \) for all \( l > L_2 \) for some \( L_2 \).

[B3] \( \beta_1(s,t) \) and \( \beta_2(s,t) \) are absolutely integrable functions.
Then, as \( n \to \infty \), \( \left| \tilde{Y}^*(t) - E[Y^*(t)|X^*_1(\cdot), X^*_2(\cdot)] \right| \xrightarrow{p} 0. \)

**Proof.** Note that

\[
f^*(t) \equiv E[Y^*(t)|X^*_1(\cdot), X^*_2(\cdot)] = \mu_Y(t) + \sum_{i=1}^{L_1} \zeta_{i1}^* P_{i1}(t) + \sum_{i=1}^{L_2} \zeta_{i2}^* P_{i2}(t), \tag{B.12}
\]

where \( P_{i1}(t) = \int_0^\Delta \beta_1(s,t)\psi_{i1}(t-s) \, ds \) and \( P_{2i}(t) \) is defined similarly. Then, we have that

\[
|\tilde{Y}^*(t) - f^*(t)| \leq |\tilde{\mu}_Y(t) - \mu_Y(t)| + \sum_{i=1}^{L_1} |\tilde{\zeta}_{i1}^* P_{i1}(t) - \zeta_{i1}^* P_{i1}(t)| + \sum_{i=1}^{L_2} |\tilde{\zeta}_{i2}^* P_{i2}(t) - \zeta_{i2}^* P_{i2}(t)|. \tag{B.13}
\]

Note that consistency of \( \tilde{\mu}_Y(t) \) follows from assumption [A1], consistency of \( \tilde{\zeta}_{i1}^* \) follows from assumption [B1], and consistency of \( \tilde{\beta}_1(t) \) follows from uniform consistency of \( \tilde{\beta}_1(s,t) \) by Proposition 1 and uniform consistency of \( \tilde{\psi}_{i1}(t) \) by assumption [B1], similar observations are true for \( \tilde{\zeta}_{i2}^* \) and \( \tilde{\beta}_2(t) \). Hence, \( |\tilde{Y}^*(t) - f^*(t)| \xrightarrow{a.s} 0 \) as \( n \to \infty \). \( \square \)

### B.5 Asymptotic Approximations for **PW** Method

From equation (4.5) we have that,

\[
\frac{1}{n} Z_j^\top Z_j = \begin{bmatrix}
\frac{1}{n} \sum_{i=1}^n \bar{X}_{11,i}(t_j) \bar{X}_{11,i}(t_j) & \cdots & \frac{1}{n} \sum_{i=1}^n \bar{X}_{11,i}(t_j) \bar{X}_{2K,i}(t_j) \\
\vdots & \ddots & \vdots \\
\frac{1}{n} \sum_{i=1}^n \bar{X}_{2K,i}(t_j) \bar{X}_{11,i}(t_j) & \cdots & \frac{1}{n} \sum_{i=1}^n \bar{X}_{2K,i}(t_j) \bar{X}_{2K,i}(t_j)
\end{bmatrix}. \tag{B.14}
\]

By the strong law of large numbers, since \( E[\bar{X}_{1k,i}(t_j)] = 0 \) and \( E[\bar{X}_{2k,i}(t_j)] = 0 \), for independently sampled data we know \( \frac{1}{n} \sum_{i=1}^n \bar{X}_{pk,i}(t_j) \bar{X}_{ql,i}(t_j) \) converges to \( \text{cov}(\bar{X}_{pk}(t_j), \bar{X}_{ql}(t_j)) \) almost surely for \( p, q \in \{1, 2\} \), and similarly, \( \frac{1}{n} \sum_{i=1}^n \bar{X}_{pk,i}(t_j) (Y_{ij} - \mu_Y(t_j)) \) converges almost surely to \( \text{cov}(\bar{X}_{pk}(t_j), Y(t_j)) \).

### B.6 Algorithms to Implement **PW** and **GB** Approaches

#### B.6.1 **PW** Implementation

Algorithm 1 shows the implementation of fitting the lagFLM model using the **PW** method based on estimation of the required smooth covariances and truncation of the KL expansion of the functional predictors. For simplicity, it is assumed that \( \lambda_1 = \lambda_2 = \lambda \) and the optimal value for regularization is chosen via cross-validation.

Inputs for this algorithm include the response \( Y_i \), the functional covariates \( X_{pi} \), and a choice for \( \Delta \). The range of values \( I_\lambda \) for regularization of the problem was chosen empirically to be an appropriate range depending on the analysis of the real or simulated data.
Algorithm 1 PW Model Fitting

1: Compute set of basis function \( \{B_{sp,k}\}_{k=1}^{K} \)
2: Estimate smooth means \( (\hat{\mu_Y}(t) \text{ and } \hat{\mu_p}(t)) \), covariances \( (\hat{G}_{X_pX_q} \text{ and } \hat{G}_{X_pY}) \), eigenfunctions \( (\psi_{pm}) \) and principal component scores \( (\hat{\zeta}_{pm,i}) \)
3: Choose values \( L_1 \) and \( L_2 \) that correspond to number of eigenfunctions that explain 99% of the functional covariance
4: Predictors are standarized by making their functional mean equal to 0 and the pointwise variance equal to 1
5: Let \( \hat{\beta}_0(t) = \hat{\mu_Y}(t) \)
6: for \( \lambda \in I_\lambda \) do
7: for \( t \in I_t \) do
8: Compute \( \hat{G}_{11}, \hat{G}_{22}, \hat{G}_{12}, \hat{G}_{1Y}, \hat{G}_{2Y}, \hat{G}, \text{ and } \hat{G}_Y \).
9: Let \( \hat{\alpha}_\lambda(t) = (\hat{G}(t) + \lambda Q)^{-1}\hat{G}_Y(t) \), where \( Q \) is the \( K \times K \) identity matrix.
10: Compute \( \hat{P}_{pm}(t) \)
11: Compute \( \hat{\beta}_p,\lambda(s,t) = \sum_{k=1}^{K} \hat{\alpha}_{pk,\lambda}(t)B_{sp,k}(s) \)
12: Compute the estimate \( \hat{Y}_i(t) = \hat{\beta}_0(t) + \sum_{l=1}^{L_1} \hat{\zeta}_{1l,i} \hat{P}_{1l}(t) + \sum_{l=1}^{L_2} \hat{\zeta}_{2l,i} \hat{P}_{2l}(t) \)
13: Compute \( NPE_\lambda = (1/n) \sum_{i=1}^{n} \frac{\{(1/m) \sum_{j=1}^{m} |\hat{Y}_{ij} - Y_{ij}|\}}{\{(1/m) \sum_{j=1}^{m} |Y_{ij}|\}} \)
14: Let \( \hat{\beta}_0(t) \) and \( \hat{\beta}_p(s,t) \) be the estimates associated to the \( \lambda \) value with smallest \( NPE \)

B.6.2 GB Implementation

Algorithm 2 illustrates how to fit the \textit{lag}FLM model using the GB method. Input for this algorithm include the response \( Y_i \), the functional covariates \( X_{pi} \), and lag parameter \( \Delta_F \).

Algorithm 2 GB Model Fitting

1: Compute smoothed predictors (\( \tilde{X}_{1i} \) and \( \tilde{X}_{2i} \))
2: Compute \( t_0, \ t \) and \( s \) matrices
3: Compute \( DX_1 \) and \( DX_2 \) matrices
4: Compute \( Y \) vector
5: \( M \leftarrow \text{gam}(Y \sim s(t_0) + s(t, s, by = DX_1) + s(t, s, by = DX_2), method = "REML") \)
6: Build \( \hat{\beta}_0(t) \), \( \hat{\beta}_1(s,t) \) and \( \hat{\beta}_2(s,t) \) from \( M \)

The smoothing of the sparse functional variables in step 1 of the algorithm can easily be performed using the methodology in (Di et al., 2009), with the \textit{fpca.sc} function of the refund package in R. The smoothed functions are sampled over \( T \) densely at times \( \{t_l\}_{l=1}^{m} \). We also define an equally spaced sampling over \([0, \Delta_F]\) at times \( \{s_{\alpha}\}_{\alpha=1}^{n} \), where we define \( \alpha \) as the index for which \( t_\alpha = \Delta_F \). Hence, for step 2, we define \( t_0 \) as the vector resulting from stacking \( n \) copies
of the vector \([t_\alpha, t_{\alpha+1}, \cdots, t_m]^{\top}\), where \(n\) is the total number of observed curves. The matrix \(\mathbf{t}\) is formed by column-binding \(\alpha\) copies of \(\mathbf{t}_0\), and \(\mathbf{s}\) is formed by row-binding \(nm\) copies of the row vector \([s_1, s_2, \cdots, s_\alpha]\). That is,

\[
\mathbf{t} = \begin{bmatrix}
t_\alpha & t_\alpha & \cdots & t_\alpha \\
t_{\alpha+1} & t_{\alpha+1} & \cdots & t_{\alpha+1} \\
\vdots & \vdots & \ddots & \vdots \\
t_m & t_m & \cdots & t_m \\
\vdots & \vdots & \ddots & \vdots \\
t_\alpha & t_\alpha & \cdots & t_\alpha \\
\vdots & \vdots & \ddots & \vdots \\
t_m & t_m & \cdots & t_m
\end{bmatrix}
\quad \text{and} \quad
\mathbf{s} = \begin{bmatrix}
s_1 & s_2 & \cdots & s_\alpha \\
s_1 & s_2 & \cdots & s_\alpha \\
\vdots & \vdots & \ddots & \vdots \\
s_1 & s_2 & \cdots & s_\alpha
\end{bmatrix}.
\tag{B.15}
\]

For step 3, the \(\mathbf{DX}_p\) matrices are defined as

\[
\mathbf{DX}_p = (t_2 - t_1) \cdot \begin{bmatrix}
X_{p1}(t_\alpha - s_1) & X_{p1}(t_\alpha - s_2) & \cdots & X_{p1}(t_\alpha - s_\alpha) \\
X_{p1}(t_{\alpha+1} - s_1) & X_{p1}(t_{\alpha+1} - s_2) & \cdots & X_{p1}(t_{\alpha+1} - s_\alpha) \\
\vdots & \vdots & \ddots & \vdots \\
X_{p1}(t_m - s_1) & X_{p1}(t_m - s_2) & \cdots & X_{p1}(t_m - s_\alpha) \\
X_{p2}(t_\alpha - s_1) & X_{p2}(t_\alpha - s_2) & \cdots & X_{p2}(t_\alpha - s_\alpha) \\
\vdots & \vdots & \ddots & \vdots \\
X_{p2}(t_m - s_1) & X_{p2}(t_m - s_2) & \cdots & X_{p2}(t_m - s_\alpha) \\
\vdots & \vdots & \ddots & \vdots \\
X_{pm}(t_{\alpha} - s_1) & X_{pm}(t_{\alpha} - s_2) & \cdots & X_{pm}(t_{\alpha} - s_\alpha) \\
\vdots & \vdots & \ddots & \vdots \\
X_{pm}(t_m - s_1) & X_{pm}(t_m - s_2) & \cdots & X_{pm}(t_m - s_\alpha)
\end{bmatrix},
\tag{B.16}
\]

which are shifted versions of the original data multiplied by the step size of the time observations. The scaling factor \((t_2 - t_1)\) is needed to account for numerical integration when performing the fitting. Note that each entry in the matrix \(\mathbf{DX}_p\) corresponds to the time associated with the difference of the corresponding entries of the matrices \(\mathbf{t}\) and \(\mathbf{s}\). Also, the first \(m - \alpha + 1\) rows are constructed by shifting the values for the first observed curve. Subsequent blocks in the matrix also follow the same pattern. This is a unique pattern observed due to the nature of our lagFLM model. The vector \(\mathbf{Y}\) in step 4 is defined as the vector of length \(n(m - \alpha + 1)\), \([Y_1(t_\alpha), Y_1(t_{\alpha+1}), \cdots, Y_1(t_m), Y_2(t_1), \cdots, Y_n(t_m)]^{\top}\), which is the vector of all stacked observations starting at time \(\Delta_F\). If the response is observed on a sparse grid of points, then we keep only the rows of \(\mathbf{Y}\) and \(\mathbf{DX}_p\) that correspond to the observation points.

Step 5 uses the \textit{gam} function in the \textit{mgcv} package in \textit{R}. The expression \(s(\mathbf{t}_0)\) fits a penalized univariate thin-plate spline at the grid points specified by \(\mathbf{t}_0\) (Ivanescu et al., 2014). The default value for \(K_0\) is 10. The expression \(s(\mathbf{t}, \mathbf{s}, by = \mathbf{DX}_p)\) is building \(\sum_l \hat{X}_{pl}(t)b_{pl}\) by fitting bivariate thin-plane spline basis \(B_{p,l}(s,t)\) at the grid points specified by \(\mathbf{t}\) and \(\mathbf{s}\). The default value for \(K\) is 30. The default thin plate spline penalty term from \textit{gam} is used.
B.7 Data Analysis using One Predictor

Here we present the results of the MTR analysis when only one predictor is used in each of the models presented in Chapter 4.

<table>
<thead>
<tr>
<th>Model</th>
<th>Predictors</th>
<th>$MSE \ [SE] \cdot (10^3)$</th>
<th>$MSE/s_0^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td>FLAIR</td>
<td>19.0 [4.2]</td>
<td>6.2</td>
</tr>
<tr>
<td>CS</td>
<td>T1w</td>
<td>7.3 [2.0]</td>
<td>2.4</td>
</tr>
<tr>
<td>CSNL</td>
<td>FLAIR</td>
<td>13.0 [3.0]</td>
<td>4.2</td>
</tr>
<tr>
<td>CSNL</td>
<td>T1w</td>
<td>7.4 [2.1]</td>
<td>2.4</td>
</tr>
<tr>
<td>lagFLM (GB)</td>
<td>FLAIR</td>
<td>5.4 [2.0]</td>
<td>1.7</td>
</tr>
<tr>
<td>lagFLM (GB)</td>
<td>T1w</td>
<td>5.5 [1.8]</td>
<td>1.8</td>
</tr>
<tr>
<td>lagFLM (PW)</td>
<td>FLAIR</td>
<td>5.5 [2.0]</td>
<td>1.8</td>
</tr>
<tr>
<td>lagFLM (PW)</td>
<td>T1w</td>
<td>5.5 [1.9]</td>
<td>1.8</td>
</tr>
</tbody>
</table>

Table B.25: Prediction error for MTR data analysis with one predictor from 10-fold cross validation. Integrated mean squared error (MSE) times $(10^3)$ along with the average $MSE/s_0^2$, where $s_0^2$ is the estimated measurement error in the observed MTR voxel trajectories.